

DOCUMENT RESUME

ED 124 426

HE 015 752

AUTHOR Young, Robert E.
TITLE Instructional Development Ways and Means. Testing and Grading.
INSTITUTION North Dakota Univ., Grand Forks. Office of Instructional Development.
PUB DATE 82
NOTE 92p.; For related documents, see HE 015 749-753.
AVAILABLE FROM Office of Instructional Development, Box 8161, University Station, Grand Forks, ND 58202.
PUB TYPE Guides - Classroom Use - Guides (For Teachers) (052) -- Journal Articles (080)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *College Instruction; College Students; Criterion Referenced Tests; *Educational Testing; Grades (Scholastic); *Grading; Higher Education; Norm Referenced Tests; *Student Evaluation; *Teacher Made Tests; Teaching Guides; *Test Construction; Test Items; Test Reliability; Test Validity

ABSTRACT

Perspectives on testing and grading are considered; and strategies, reference materials, and sample test items are presented. Attention is directed to specific impacts of testing and grading on students and teachers, purposes of testing and grading, and common complaints of students and teachers. Extra activities proposed for the teacher include discussing with colleagues the testing/grading purposes and keeping a record of student complaints regarding grading and tests. Several prerequisites are identified as important for test development: subject matter expertise; well-developed instructional objectives; mastery of written communication; an understanding of the individuals for whom the test is intended; and knowledge of the principles and procedures of item writing/test construction. A 15-step strategy for planning, implementing, and evaluating classroom tests is presented, based on test construction principles. Extra learning activities are proposed for the teacher to provide additional information about testing principles and practices, practice with various aspects of test construction, and critique of the teacher's tests and testing procedures. Several purposes of grading (reporting; decision making; and stimulating, directing, and rewarding students), and limitations of grading and norm- and criterion-referenced tests are considered. A bibliography; information on score distributions; examples of types of test items; and articles on test reliability and validity, tables of specifications, and norm- and criterion-referenced tests are appended. (SW)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED224426

INSTRUCTIONAL DEVELOPMENT
WAYS AND MEANS

Testing and Grading

Robert E. Young, Ph.D.
University of North Dakota

Requests for copies
should be made to
Office of Instructional Development
Box 8161, University Station
Grand Forks, ND 58202

Copyright
Robert E. Young
1982

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

This document has been reproduced as
received from the person or organization
originating it.

Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official NIE
position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

Robert E. Young

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Testing and Grading: A Means to Effective Teaching

Effective teaching depends significantly on the fair and appropriate use of tests and grades. Tests serve a number of purposes. They assess, educate, and motivate. Grades have their own purposes, as well. They represent teachers' responsibility to judge and report students' educational performance. Testing and grading, despite their importance, can be the most difficult and least rewarding part of a college teacher's job. Creating tests and assigning grades can become less problematic and more useful to the teaching and learning process, though, if you keep in mind the nature of testing and grading and consider a small number of principles and procedures. This package has been designed to provide a better understanding of testing and grading and to offer ways to become more skilled at testing and grading.

These materials can be used independently or with a group of colleagues. In the three sections that follow you will find a presentation of ideas, strategies, and examples. Each section has "text" material, followed by objectives for the section. Extra Learning Activities, following each section, provide opportunities for active thinking, discussing, and testing of the ideas and approaches presented.

CONTENTS

- I. The World of Testing and Grading--Impacts, Purposes, Problems
- II. The Technology of Testing--Principles and Practices
- III. Grading--Philosophies
- Appendix

SECTION I

The World of Testing and Grading--Impacts, Purposes and Problems

Teaching is a "real world" activity, despite comments to the contrary. You need only think of the impacts testing and grading have on the lives of students and teachers to know that teaching is serious business.

Students expend physical, mental and emotional energy preparing for tests, quizzes, and other evaluation activities. They usually exhibit anxiety when approaching an exam. Some even expose themselves to physical distress while preparing to be evaluated in a course. Test results have definite impacts on students' lives as well. Test results can make students feel happy or sad, successful or inadequate, rewarded or unfulfilled. More importantly, test scores can propel students toward realizing their life goals, or they can defeat those ambitions.

Testing and grading affect teachers' lives in similar ways. Preparing fair, yet challenging exams requires hard work. Grading--the assignment of a certain value to a student's performance--can be an excruciating process. Successful test performance and high grades bring joy and a sense of accomplishment to both teacher and student. Poor performance and low grades often carry disappointment, a sense of failure, and even a sense of guilt.

If testing and grading is serious business--real world stuff--for both students and teachers, it is also important business. Any activity which commands so much energy and emotion must be important. Over one

hundred million classroom tests are given in higher education each academic year (Milton and Edgerly 1976). There must be some essential purpose served by testing and grading.

Though often obscured, testing and grading do have important and essential purposes. Students depend on testing and grading, in part, to inform themselves and others about what and how much they know. Tests and grades inform students where they need to know more. Tests and grades motivate students and influence them to study in certain ways.

Testing allows a teacher to find out if the students have learned that which he wanted them to learn. Tests scores tell him which of his students need additional attention. Tests also tell him whether his teaching has been effective. Finally, tests and grades allow the teacher to fulfill his responsibility of informing students, future teachers, and employers of accomplishments in his course.

This listing of purposes suggests that testing and grading have a very real and important role in the educational process. Tests and grades help learners learn, and teachers teach. In fact, without tests and grades learning and teaching would be less effective and less meaningful.

Testing and grading, though, are not without problems. Much of the energy focused on testing by students and teachers has to do with correcting or guarding against unfair and inappropriate testing and grading practices. Students on some campuses spend considerable time trying to protect themselves from questionable practices. A survey of university ombudsmen indicates that nearly thirty percent of all student and faculty complaints had to do with testing and grading. At some schools, the figure approaches eighty percent (Milton and Edgerly 1976).

Some of the specific complaints included:

1. In an introductory course covering approximately 2,500 years of philosophy, five of the seven questions on the final exam were about Kant. (Good testing should be based on representative sampling. Five of seven questions focusing on one philosopher within a period of 2,500 years is not adequate sampling).
2. Five textbooks were assigned in a course designed to promote understanding of the contributions of several individuals in the field being studied. Lectures and class discussions focused almost entirely on one of these people. All questions on the final concerned the one person. (Instructors need to be much more careful than many seem to be in correlating objectives, assignments, and testing.)
3. Students in a senior course were assigned 15 journal articles, the shortest of which was 14 pages long. The only examination question over this considerable volume of material asked students to match the articles' authors with the titles. Challenged by a colleague, the instructor argued he could assume that a student who could do this matching understood the material. (The assumption made by this professor requires a major leap in logic.) Moreover, his future students might play the odds and limit their learning to memorizing authors and titles.
4. Students in a technology course were assigned a project to be completed individually. They were informed that the only grading criterion would be the quality of the project. Several students devoted many hours to the assignment early in the term and finished several weeks before final exams. Other students maintained a leisurely pace and worked throughout the term. All the early finishers received F's, while the leisurely ones received A's and B's. The instructor maintained that many of the F's were awarded because of absences--he had not seen some of these students during the last month of the course. Even when reminded that his announced criterion was quality, not class attendance, he refused to alter the F's. (Many teachers are unduly touchy about class attendance. To be fair and just, measurement should measure what has been asked for and nothing else.)
5. After assigning quite high mid-term grades, an instructor declared that students were being coddled. Without any warning, term papers were graded very harshly (after the withdrawal date had passed). The final was graded equally severely. Course grades for the class of over 40 included one A, one C, and a few D's; the rest were F's. Colleagues

of the instructor arranged for the students to take another final. (Capriciousness and arbitrariness have no place in evaluation.)

6. A freshman was told by an instructor she would receive a B in his course; her final grade was C. The student was applying for admission to a competitive program for which a few hundredths of a point in her GPA might determine her acceptance. Investigation revealed that the instructor was a teaching assistant who had left school. The department chairman believed this instructor's teaching and testing methods had been questionable and changed the grade. The student was admitted into the competitive program. (Most teaching assistants do not receive formal instruction or guidance about testing and grading. Incidents such as this one may be prevalent.)
7. A freshman who had maintained a C average on all his tests received a final grade of F. The instructor explained that the student had exhibited an improper spirit toward the subject matter and refused to alter the grade. (We have serious doubts about the propriety of grading a student's "spirit" or attitude.)
8. A final was administered to 130 students in a crowded classroom where it was easy for students to copy from each others' papers. Although many students thought the situation was unjust, the instructor refused to change the test location, blaming the institution for assigning too many students to the classroom and for providing the small classroom. An administrator, the department chairman, and the college dean intervened, and the test was given again under satisfactory conditions. (Inadequate and improper testing conditions are inexcusable.)

These complaints and others (you may have some yourself) will continue to be lodged. Testing and grading are emotional issues; much is at stake whenever a teacher prepares a test, a student takes it, and the teacher reports the result and gives a grade. There are things, though, that can be done to minimize the complaints, accomplish the purposes of testing, and maximize the positive impacts of testing and grading on both students and teachers. One of these things happens to be what you are doing right now: reading, studying, and thinking about testing and grading of students.

Objectives of Section I

From your reading of this section, you should be able to describe some of the impacts which testing and grading have on students and teachers.

Impacts of Testing and Grading on Students and Teachers

- causes the expending of considerable physical, mental and emotional energy (students and teachers)
- causes anxiety (students and teachers)
- can produce physical distress (students)
- produces feelings of joy, happiness, accomplishment, success, and reward or feelings of sadness, disappointment, inadequacy, and failure and guilt. (students and teachers)
- advances or hinders one's ambitions (students and teachers)

You also should be able to list the essential purposes of testing and grading.

Purposes of Testing and Grading

- tells students what they know and how well they know it.
- points students to areas where they might need or want to know more
- motivates students
- influences students
- influences how students study and what they learn
- tells a teacher if his students have learned that which he wants them to learn
- tells a teacher which students need more attention and in what areas
- tells a teacher whether his teaching has been effective
- allows a teacher to communicate the nature of the student's performance to others (one of his responsibilities)

And, you should be able to state some of the common complaints which students and teachers have about testing and grading.

Common Complaints

- tests do not cover a fair representation of course material
- tests do not follow from course objectives or content
- tests ask for trivial information and do not assess more important understandings
- grades are sometimes based on things students were not told about in advance (e.g., attendance)
- teachers are capricious and arbitrary in assigning grades--many teachers have no criteria to guide or justify their grading
- teacher's grade on qualities that are impossible to measure accurately (i.e., spirit or attitude)
- some teachers allow social biases (i.e., against women students) to influence their grading
- testing conditions are often poor (e.g., crowded or uncomfortable, or noisy rooms)
- complaints of your own or others

Finally after thinking about what you have read (and possibly after doing some of the things suggested below as "Extra Learning Activities") you should be able to 1) describe to your students the purpose of testing and grading in your courses, 2) identify complaints which your students have about your testing and 3) recognize and be able to list the impacts which your testing and grading has on you and your students. After completing the rest of this package, you will also be able to eliminate these complaints and maximize the positive impacts of testing and grading on you and your students.

EXTRA LEARNING ACTIVITIES

These activities have been designed as ways of gaining additional understanding of the purposes and impacts of testing and grading and the complaints which students and faculty have about these activities. Each contributes to the objectives of this section, and they have been designed to assist you to be even more effective in testing and grading your students, an important aspect of effective teaching.

1. Purposes of Testing and Grading Your Courses

a. Testing and grading in your courses will have definite purposes. It is probably important that you consider what these are. You can refer to written descriptions of your courses, but that may not tell you all you want to know. To better understand these purposes, arrange an individual or group discussion with other members of your department to discuss the purposes of testing and grading in your courses, and in the department, college, or university.

b. Since it is important that students understand the purposes of testing and grading in their courses, make a plan to communicate these purposes to them. A verbal presentation and/or a written handout might be ways of doing it. A brief discussion of these purposes with students will allow you to determine if they are understood.

2. Testing and Grading Complaints

a. Students may have complaints about the way you test and grade them, despite your best efforts. As a way to understand these complaints and as a basis for eliminating them, you might simply record any complaints which

students have about what you do in this area of your teaching. A pattern of complaints during one term or over a number of terms might indicate an area for some work on your part. Sharing of lists with other teachers might point out areas for the entire group to consider or a place where the program might need some change. In addition to recording student's complaints, you may identify your own concerns and those of your teaching colleagues. You might also identify complaints which other faculty have about your testing and grading.

The remaining sections of this package will present principles and practices by which most common complaints can be eliminated. The first step though is to carefully identify them.

3. Impacts of Testing and Grading

Identifying the impacts which testing and grading have upon you and your students will be enlightening in and of itself, but it might also suggest ways in which your testing and grading practices are succeeding and/or failing. Try to recognize and make a list of the impacts that your testing and grading has on students and yourself. Observe and/or ask students about their feelings and actions at the beginning of the course, before a test, when they get the results, and at the end of the course. Record your own feelings and actions as you think about and prepare your tests and grades.

Again principles and practices presented in the next section will suggest ways to minimize negative impacts and maximize positive impacts.

Section II - Testing

Testing, like teaching in general, is a creative and artistic endeavor. Once viewed this way much of the drudgery of testing can be avoided.

The problem to be solved by you as tester-artist is this: How do I determine what students have learned in a way that will give me, my students, and those to whom I report confidence in those results?

To solve this problem you will have to rely on a range of abilities. Whether you are preparing a lengthy major course examination or a short weekly quiz, your knowledge of your subject matter, your ability to express yourself, and certain principles and procedures of test construction will be important in developing fair and appropriate tests of your students' understanding.

In this section you will be presented: 1) a list of abilities necessary for successful test construction. You will also be provided a planning strategy for using these abilities to design fair and appropriate tests.

Abilities

Any art is built on a set of abilities. The same is true with testing. The basic abilities required in testing and grading can be learned. Many have already been mastered, and the others can be acquired through the kind of information and practice offered in this section.

The creative synthesis of all these abilities is another matter though. It too can be learned, but in quite a different way. Practice is

important. But the essential component becomes the careful critique by other tester-artists. Extensive and critically supervised practice is an essential ingredient of becoming a successful tester. This practice and critique results in high standards of quality and pride in craftsmanship.

This, though, is getting ahead of the story. First, the basic abilities themselves.

A number of testing experts have ventured to describe the abilities necessary to produce good tests. (Wessman, 1969; Milton and Edgerly, 1977; Eble, 1972). The following list represents a distillation of these views. To be a competent test constructor, you need to possess at least the following characteristics.

1. Subject matter expertise. You must not only have a knowledge of the facts and principles in your subject, but you must thoroughly understand them. You need to know the implications of basic ideas and need to be aware of the popular misconceptions and fallacies in the field. This is particularly important when designing structured-response items such as multiple choice or true-false, where false statements are used in designing items.

2. A well-developed set of aims or objectives for instruction. This characteristic involves both broadly conceived educational values and specific course outcomes. Educational values will determine, in large part, how you approach your students and your responsibilities as a teacher. Educational values will determine, in large part, how you approach your students and your responsibilities as a teacher. Educational values have already been expressed in this package. For example: testing should be

aimed primarily at helping students learn. If you adhere to this value, your testing format and procedures will be quite different than if you believe testing's only function is to certify how much a student knows about your subject.

In addition to educational values, you need to be clear about the specific knowledge, attitudes, and skill outcomes you desire from each course you teach. These outcomes are usually expressed in the form of course objectives.

Well-stated objectives provide the basis for any test or quiz. Once objectives are specified, the purpose of testing becomes one of determining whether students have or have not accomplished the objectives.

Test questions flow directly from the language of the objectives. For example, in a Biology course the following objective appears: ". . . make a diagram of the cell cycle and describe all the events that occur." An appropriate test of this objective would ask students, "Make (draw) a diagram of the cell cycle and describe each of the events that occur in the cycle."

In some courses, the course objectives have already been specified. Not having to do this yourself makes your job somewhat easier. As you will see later in this section, planning a quiz or test begins with the objectives as stated. Other courses which you teach in the future may not have objectives specified for you in advance. You will have to do this as part of the planning for the course. In the Appendix (Handout 2) to the package, you'll find a list of references which will help you think about and prepare instructional objectives.

3. Mastery of Written Communication. Expressing exactly what you intend to say in a way that students can easily understand is essential in giving an effective test. The listing of this ability may seem obvious and even trivial. But many tests become ineffective because students legitimately do not understand what the teacher is asking.

Words must be used precisely and arranged in ways to effectively yet simply, communicate what is intended. An awareness of possible misinterpretations will guide effective test item writing.

4. An understanding of the individuals for whom the test is intended. Students at different levels of education have different levels of educational development and characteristically different ways of mental processing. Students at any one level, such as freshman-level History, will also differ among themselves on how much they know about the subject before the class begins, and how they might learn best.

An understanding of your student's level of educational development will allow you to prepare tests which challenge them yet do not confuse and intimidate them.

5. Knowledge of the principles and procedures of item writing/test construction. A working knowledge of principles for planning a test, different types and varieties of test items and their advantages and disadvantages, characteristics of good items, and common errors will allow you to create fair and effective tests. Although testing has been characterized as an art, there is some science which can be applied to the process. Research on various aspects of testing suggest principles and effective practices.

The next sub-section will present these principles and principles of test construction as part of a strategy for classroom testing.

A Strategy for Planning, Implementation and Evaluation of Classroom Tests

A successful test constructor has a strategy or set of strategies by which he approaches the task of preparing an examination. Below is one strategy which reflects the best thinking of testing experts. This strategy is based on a set of principles of test construction. These principles will be offered at each step of the strategy.

The strategy as a whole involves 15 steps. It can be adapted specifically to the testing situation which you face in your subject area.

1. Identify the content and/or objectives to be tested.
2. Determine the type or types of test items to be used.
3. Decide which objectives to test.
4. Prepare test items for each objective.
5. Group items in the test according to their type and arrange the items within each group in descending order of difficulty.
6. Randomize the response patterns for true-false, multiple choice and matching items.
7. Prepare instructions and test directions.
8. Pilot-test the test.
9. Revise items, test, and instructions where necessary.
10. Prepare final test and scoring key.
11. Administer the test.
12. Score the exam.
13. Produce relevant test statistics.
14. Analyze the items.
15. Derive grades.

1. Identify the content and/or objectives to be tested.

Principle: Tests should be based only on the stated objectives and content of the topic or unit in question.

As stated earlier, once objectives of a learning experience have been specified, the purpose of testing becomes one of determining whether students have not accomplished the objectives. Often, though, test questions ask for information or skills not specified as intended outcomes of the unit or course being tested. For example, if the objective suggests that the student should be able to "list four important chemical functional groups", then it is appropriate to ask him to "draw the structure of four important chemical functional groups and describe the unique chemical properties of each" on the test. Such a question requires a different ability than that called for by the objective and even though it may be related, the student was unaware that he was to acquire the ability and the learning experience may have been such that he did not have an opportunity to learn it.

2. Determine the type or types of test items to be used.

Principle: The type of test item or question used should be appropriate to what is being measured.

A wide variety of test items is available to the college teacher, including problems, essay, short answer or multiple choice, true-false, completion and matching. Problems, essay questions and short answer items do a better job of measuring the student's ability to analyze and apply and should be considered when these kinds of objectives need to be tested.

These types of questions are generally time-consuming to answer and tend to limit the amount of content sampled in the test. Completion, true-false, and multiple choice items are useful for coverage of more content, and although higher order cognitive abilities can be measured by such items, usually objectives that get at facts, concepts and principles will be assessed by these types of items. Knowing the merits and limitations of each type of question in relation to the purposes of the test should enable the teacher to select the appropriate categories of test items.

3. Decide which objectives to test. If all the objectives of a unit are important and if all have been covered in the instruction of the unit, then you ideally should test each objective. Often, though, not all objectives are important and/or some may not have gotten much instructional attention. Only test those objectives about which you want to know about student's success and accomplishment, and only test those items which you feel received coverage sufficient to allow accomplishment by a majority of the students.

Principle: The test or quiz should adequately sample the content involved.

Content should be tested in proportion to the degree to which it was emphasized in the course. Those topics which received the most attention in the unit should get a proportional amount of attention on the test. For example, if you are designing a test with 25 items and you decide that topic A is twice as important as topic B then twice as many items should be devoted to A as to B.

More typically you do not have time to test all the objectives. In this case, test those which are most important and those which received the most attention during the unit. Or, sample the number of items needed

from the different topic areas and types of objectives (facts, concepts, principles, problem-solving) according to their importance in the unit.

Not testing for all objectives does, though, present a problem. Unless you can show that ability to accomplish one objective implies an ability to accomplish others not tested, you will not be able to say whether the non-tested objectives have been achieved or not.

4. Prepare test items for each objective. Prepare at least one item for each objective. Write each item on a 3x5 card. Cards are a convenient way to store items for when you teach the course again. Also, the reverse side can be used to record information from item analysis (to be discussed below).

Preparing more than one item for each objective will give you important practice in item writing. Also, it will give you some choice as to which items to use on the test. You might try writing questions of several different formats for each objective.

If time and test length permits, asking an additional question for each objective increases the reliability of the test.

Principle: You can be more sure that a student actually does have the ability required by the objective if he can answer correctly more than one question. One well-developed question can, though, give you the reliability you need.

5. Group items in the test according to their type and arrange the items within each group in a descending order of difficulty. Items which require more time to complete (i.e. essay, problems) should come at the end of the exam.

Principle: Students will maintain a higher level of motivation and confidence throughout the test if they begin with easier items and proceed to more difficult items.

Also, each type of test item requires a slightly different thought process to interpret and answer. Grouping items by type minimizes the number of times students must change the general way they attack a question.

6. Randomize the response patterns for true-false, multiple choice and matching items. You want to avoid the situation where students can answer correctly by simply figuring out your pattern of answers.

Principle: Student "test-wiseness" in the absence of real ability to accomplish the objective being tested, can lead to successful test performance.

Whenever this happens, the reliability and validity of the test will be diminished.

To randomize you can use a table of random numbers.

7. Prepare instructions and test directions.

Students often do poorly on exams, not because they lack the required knowledge or skill, but rather because they were unclear about what the test asked them to do.

Principle: Providing explicit instructions and directions on the test or verbally before the test will help insure that students understand testing task(s) the way you intend it.

If students must move about in a test, if they must answer questions in a particular order, if they must choose a subset of questions to answer, or if they must answer according to a particular format or if they have a time limit, clear written or verbal instructions in advance results

in a smoother and less confusing test for your students and you.

"Tips" for how to approach the tests can also be useful, particularly for students who experience anxiety in your testing situation.

8. Pilot-test the test. More often than not this step gets left out in test construction. Time pressures usually do not allow this check on the quality of a test. This step, though, can contribute more than any other to effective tests and quizzes.

Principle: An ounce of prevention (for unfair or unclear items) is worth a pound of cure (for student anger or disappointment)

Ask your colleagues, at this University or elsewhere, to review and critique the test for content, format, and clarity. Ask graduate students to do the same; your research or teaching assistants may provide you with the best critique. Ask former students (with appropriate test security precautions, of course) for comments on clarity of items and instructions. Ask your wife/husband or girlfriend/boyfriend if they understand what you are asking. Ask testing experts. OID staff will critique tests with the view toward helping you improve your testing ability.

9. Revise items, test, and instructions where necessary. Use the comments received to improve the items. Make corrections on the 3x5 cards or prepare a new card.

10. Prepare final test and scoring key. Tests must be put in a form which can be given to a student (i.e., newsprint, chalkboard, projector, duplicated copies). Also, the scoring key should be prepared in advance of giving the test. For structured-response test, preparing a key means simply deciding the correct response for each item. For short answer or

essay tests, a model answer should be prepared. A model answer identifies criteria against which students' work will be assessed.

Principle: Knowing specifically what an answer should include before reading a set of tests, and determining the point or grade value of varying degrees of satisfactory performance will increase the objectivity hence the reliability of a test.

11. Administer the test. Check the physical conditions of the room to insure adequate light, proper temperature, and humidity. Arrange seating to reduce the temptation for cheating. Carefully monitor the exam without appearing to distrust students.

Principle: Physical, social, and interpersonal conditions should be such that they do not detract from student's opportunity to display their true abilities.

At the very least, conditions should be similar for all students.

12. Score the exam. It has been mentioned previously that preparing in advance a scoring key or model answer makes scoring a straightforward matter of comparing students' answers against the key or model answer and assigning the pre-determined number of points to each answer.

In the case of quizzes and tests of the short answer or essay variety, you need to grade answers yourself.

Principle: Fatigue on the part of the grader can often result in students' scored last being evaluated differently than those scored first.

This is not usually a problem for short quizzes, but on longer tests short breaks throughout your scoring process can eliminate this fatigue effect. Also, scoring sometimes can be more effective if you score one question for each student before moving on through the remainder of the

test. It is easier to keep one model answer in mind at a time.

13. Produce relevant test statistics. Test score statistics include frequency distributions; mean, median, and mode, variance or standard deviation, percentile and standard scores. Since you will be grading on an absolute scale (i.e., 100-93 = A etc.), it is not essential that you determine any of these statistics for reporting and grading purposes. You and students, though, may find interesting the distribution of scores, the average or some other statistical description of scores. A handout in the Appendix (Handout 3) will describe each of these statistics and means of computing them.

14. Analyze the items. You can analyze the effectiveness of your quiz or test in formal and/or informal ways. Formally, you can determine how effective the various items were in discriminating between those who did well and those who did poorly on the test as a whole. You also can determine a degree of difficulty for each item on the test by computing the percentage of students who failed a question.

For tests which can be answered or transferred to an optical scan scoring sheet, the UND Computer Center will provide scores and discrimination and difficulty indexes. You should request from the User Services Manager a copy of the test analysis material. These can be helpful in constructing future tests of more or less difficulty or discrimination. Also, they are helpful in identifying areas where large numbers of students are having trouble.

Informally, you can be aware of any complaints students direct toward items on the test. Or, you can routinely ask students after the exam

to point out items they felt were unfair, inappropriate, or poorly written.

15. Derive grades. Since testing and grading are considered distinct activities, done for different purposes, grading will be considered in the next section.

Suffice to say here that testing provides the basis for grading. Tests tell you how much a student knows. Grading then attaches a worth of that result. Testing, though, has important purposes apart from grading. Those purposes were detailed in Section 1 of this package.

Learning The Art of Testing

Early in this section, a necessary synthesis of abilities into an art of testing was mentioned. The abilities and their integration can be mastered, it was suggested, by knowledge, practice and critique. Preparing tests and test items and having them scrutinized by colleagues, testing experts, students, and yourself will be the surest way to becoming a successful tester-artist. Designing tests to be effective at facilitating students' learning and assisting you in your job of assessing and reporting student progress requires some competency at the abilities described and a willingness to learn from your mistakes. Competence and artfulness at testing does not develop overnight. It's a learning process.

The extra learning activities for this section offer ways of developing the requisite abilities and of practicing and getting your tests critiqued.

Objectives of Section II

From your reading of this section, you should be able to list the basic abilities necessary for successful test construction.

Abilities Necessary in Testing

- subject matter expertise
- well developed aims and objectives for instruction
- mastery of written communication
- understanding of students capabilities
- knowledge of principles and procedures of item/test construction

You also should be able to list and describe the steps involved in planning, implementing, and evaluating classroom tests.

Strategy for Test Construction

1. Identify the content and/or objectives to be tested.
2. Determine the type or types of test items to be used.
3. Decide which objectives to test.
4. Prepare test items for each objective.
5. Group items in the test according to their type and arrange the items within each group in descending order of difficulty.
6. Randomize the response patterns for true-false, multiple choice and matching items.
7. Prepare instructions and test directions.
8. Pilot-test the test.
9. Revise items, test, and instructions where necessary.
10. Prepare final test and scoring key.
11. Administer the test.

12. Score the exam.
13. Produce relevant test statistics.
14. Analyze the items.
15. Derive grades.

In addition, you should be able to state the principles of test construction embodied by the strategy.

Principles of Test Construction

- tests should be based only on the objectives of instruction
- type of test item used should be appropriate to what is being measured
- exam should adequately sample the content covered.
- the more items used to test an objective, the more reliable the test of that objective.
- confidence and motivation will be facilitated by ordering questions from easy to hard, and concentration will be enhanced if questions are grouped according to type.
- test-wiseness, not relevant knowledge and skill, can lead to successful performance on a test, thus lowering reliability
- explicit test directions help students better understand what to do on an exam.
- review of a test before it is given can eliminate unfair and unclear questions and directions.
- preparing a scoring key or model answer in advance helps eliminate subjectivity in scoring.
- physical, social, and interpersonal conditions should not distract students when being examined.
- scorer fatigue can affect scores given to student answers and papers.

Finally, after thinking about what you have read and after doing some of the Extra Learning Activities, you should be able to 1) design and

implement a quiz or test which is valid, reliable, and useful to students' learning and your responsibilities for assessing and grading students' performance. You also should be able to 2) critique the tests of others and make suggestions for improving their effectiveness.

EXTRA LEARNING ACTIVITIES

These activities have been designed to provide you with additional information about testing principles and practices, practice with various aspects of test construction, and critique of your tests and testing procedures. Each contributes to the objectives of this section.

1. Abilities Required of a Test Constructor

a. Popular misconceptions and fallacies surrounding the basic ideas in your subject will provide good "foils" for test questions, and they will be something that you will try to overcome as you teach students. You might meet regularly with colleagues and advanced students to identify and discuss these falsehoods.

b. Try listing the educational values you hold most important. Now think how these values might affect the way you approach or should approach testing.

c. Arrange through the Writing Lab (102 Old Science) to have a sample of your test items critiqued for effective written expression by an English faculty member. You might also ask your students to review the questions you have given them so far this term and indicate to you where they had problems of understanding the questions.

d. Read other articles, books, and pamphlets on testing listed in the Bibliography.

2. Strategy for Test Construction

a. A fifteen step strategy puts a terrific burden on one's memory. Can you shorten it for your own purposes by collapsing several steps into one? Don't leave out anything important.

b. To learn more about the advantages and disadvantages of different types of test items, take a look at Handout 4 in the Appendix.

c. A "table of specifications" can be a useful way of insuring that your test adequately covers unit or course content. Consult Handout 5 in the Appendix.

d. Throughout this Section, issues of reliability and validity have been dealt with in a variety of ways. For a more formal discussion of these concepts, see Handout 6 in the Appendix.

e. For a more formal discussion of item analysis--discrimination and difficulty indices--again consult the Appendix; Handout 7.

3. Learning the Art of Testing

a. Ask one or more colleagues to critique all the tests you have written this term. By having more than one peer look at your tests you benefit from seeing how your choice of item format and written expression affects different people.

b. Form a partnership with another person to critique each others tests the remainder of the term. Do this before you give a test so the feedback you get can help you improve the test.

c. Ask someone at the Office of Instructional Development (OID) to review your tests (they will be happy to do it).

SECTION III - GRADING

Assigning grades may be the most disconcerting aspect of a teacher's job. Grading involves judging the worth of a student's performances and communicating that judgment to the student and others. A "C" on the UND grading scale technically means "average," and in this day of grade inflation really less than average. Telling a student that he is average is not the easiest thing to do, particularly if you see other outstanding qualities in the person. Giving a low grade may mean that the student's chances for certain careers will be hindered. Few teachers shoulder this responsibility with no anxiety. But grading is a responsibility of every teacher in our educational system (at least the ones we know at UND), and despite its negative aspects grading serves important purposes.

Purposes of Grading

Grades provide an overall summary index of achievement in a course. This index serves the purposes of reporting, decision-making, and stimulating, directing, and rewarding students.

The reporting function of grades is probably its most important and useful. It also is the least controversial use of grades. Grades are foremost a means of communication. They are an efficient way of informing students, parents, admissions officers, and employers and others of the nature of a student's academic achievement. Grades represent an agreed upon way of reducing a large amount of information about student performance into a form that is easily communicated and stored for future use.

Future use usually means some decision-making about a student. Should he be admitted to medical school? Should he be given a fellowship? Should he be hired into the firm? Although the evidence is mixed, grades at one educational level predict the kinds of grades a student will get at another level and, to some degree, success beyond his schooling. This relationship of grades to future performance makes grades a legitimate basis upon which to make decisions. Even where the relationship is weak, grades are and will continue to be used for decision-making. This fact makes it essential that care be taken to assign grades that truly reflect the extent of a student's educational achievement.

Grades also have a stimulating, directing, and rewarding effect on students. A high grade might encourage a student to continue his study in an area. A low grade might steer him out of an area or direct him to additional study to correct the deficiencies suggested by the grade. Grades also can be effective reward for hard work and a job well done. Again, accurate grading is important if students are to get adequate messages about their performance and potential.

Limitations

The type of grading procedures used in American higher education have definite limitations. It is difficult for one instructor or one institution to completely overcome these problems but an understanding of them will hopefully cause you to take even more care as you go about assigning grades to your students.

First, the meaning attached to each grade is not the same across institutions or even across instructors in any one institution. An "A" is not an "A" is not an "A" . . . or a "2.0" is not a "2.0" is not a

"2.0" . . . As units of measurement these symbols do not have uniform meaning, as does the inch or the pound. The definition of each point on the grading scale is not clearly defined, generally accepted, or uniformly observed. In more technical terms, grades are often invalid measures of performance.

In some programs this problem has been eliminated, in part. A standard grading scale has been developed to be used by all instructors. Given that tests accurately measure performance, students with the same level of achievement will get the same grade. An "A" is an "A" is an "A" in these programs. This does not, though, eliminate the possibility that an "A" at North Dakota State might represent higher or lower achievement.

A second problem with grading has to do with the evidence upon which grades are based. In many courses grades are based upon insufficient, irrelevant, and subjective evidence of achievement. This might mean the lack of objective measures of performance such as tests or the use of poorly constructed measures. The technical issue here is the reliability of the judgements made by teachers of student performance. Without objective, relevant, and well-designed measures, students, if they would take the course twice from the same instructor or from different instructors and achieve at the same level, may receive different grades.

This problem can be solved easily enough. If regular quizzes and tests are given by each instructor, and if instructors have the ability to prepare effective tests, then sufficient and relevant evidence should be available when it comes time to assign grades.

A third problem with grades concerns their potential detrimental side effects. The impact of failure can be debilitating. Grades can lead to excessive competitiveness, cheating, and a distortion of educational values where grades become more important than learning itself.

Finally, grades as a communication vehicle have limitations. Grades do not provide information specific enough for a student to know exactly where his strengths and weaknesses lie. They represent gross measures of performance, often satisfactory for making decisions about admission, employment, and advancement. But, they do not tell a student what he knows or how well he knows it in any specific and diagnostically useful way.

Grading Practices

One of the purposes of grading mentioned above was reporting student's level of achievement. Exactly how that level is determined and translated into a grade was not specified. This section suggests two distinct ways of doing this translation. Again, purposes become important considerations.

Basically grades can be determined by comparing performance to some absolute scale set by the teacher or by comparison with the performance of the class (or some similar group). This latter approach is often called grading "on the curve."

Relative-scale grading. College teachers typically practice relative scale grading in two ways: Natural breaks in the distribution of scores or statistical determination of the number of A, B, C, D, F grades to be given.

Natural breaks. When scores are ranked from bottom to top often the instructor will find scores clumped together along the continuum. One might use the natural breaks in the distribution as indicated in the example below, as places to differentiate among the various grade levels.

98 95 94 93 85 83 80 78 72 72 71 71
69 69 68 67 66 61 60 58 57 48 43 41

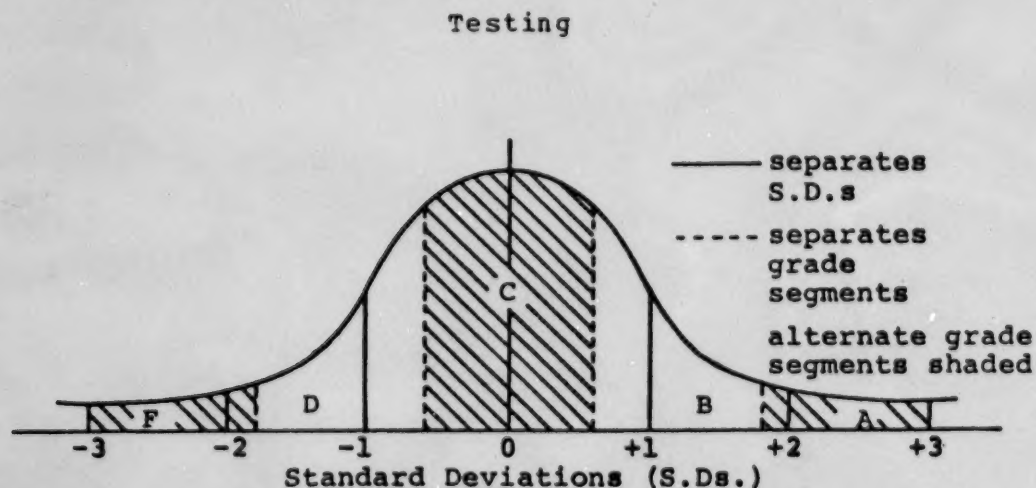
These scores can be grouped as follows:

Group 1:	98 95 94 93
Group 2:	85 83 80 78
Group 3:	72 72 71 71 69 69 68 67 66
Group 4:	61 60 58 57
Group 5:	48 43 41

Statistical. This procedure takes into account the variability of scores and assumes a normal distribution of scores. It is particularly useful where a large number of scores (100 plus) are involved.

Since essentially all of the scores will lie between plus and minus three standard deviations of the mean, we can assume that the entire distribution falls within this six standard deviation range.

For illustration purposes, let's assume a 5 point grade distribution scale. This means that the six standard deviation regions of the curve must be divided into 5 segments:



These grades must be distributed about the mean with $1/2$ of the C's or $1.2 \text{ standard deviation} / 2 = 0.6 \text{ standard deviation}$ above the mean and $0.6 \text{ standard deviation}$ below the mean. Assuming the mean is 50, and the standard deviation is 10, then $0.6 \times 10 = 6 \text{ points}$. The students with the scores from $50 - 6 = 44$ up to those with scores of $50 + 6 = 56$ will be given the same grade, in this case a C. Therefore, all scores between 44 and 56 are given a grade of C. The D's start at 43 and extend down to 1.2×10 or 12 points lower, at 32. That is, the D range is 43 to 32. The B's start at 57 and extend up to 12 points above 56 (68). The A's then start at 69 and the F's include any score below 32.

The system described above assumes a "norm referenced" grading scheme. That is, grades are based on the relationship of the individual's work as compared to the others in the same class. It is important to recognize this fact. In today's world, very few instructors would set the mean grade equal to a C. You ought to be familiar with "expected" grades on your campus before you assign grades in this manner. Also, a symmetrical distribution is likely to be skewed to the high grades.

Absolute scale grading. Grading from an absolute scale is making a resurgence in American higher education after a 40 year dominance by relative grading. As typically practiced it means developing a scale which remains the same no matter what the scores of the class on a test. Students are notified in advance that they must score correctly on a pre-determined percentage of the items in order to achieve a specified grade (e.g., $90/100\% = A$, $80/89\% = B$, etc.). This approach is referred to most frequently today as "criterion reference" grading.

Using a specified absolute scale, you might use the following procedure or some modification to assign grades to students in your classes.

1. During the term record scores for each quiz, test, or paper in a grade book.
2. At the end of a term:
 - a. determine the total possible points (e.g., multiply number of quizzes by 10 points and the number of tests by 100 points, then add these two figures)
 - b. total the number of points earned by each student (from your grade book)
 - c. determine the percentage of points each student earned out of the number possible (total points earned $\div$ total possible points $\times$ 100 = final percentage score).
3. Using the scale you developed, determine the grade equivalent of each student percentage score. This will be the grade you record for the student.

OBJECTIVES FOR SECTION III

After reading this section, you should be able to state and describe three broad purposes for grading.

Purposes for Grading

- Reporting
- Decision-Making
- Stimulating, Directing, and Rewarding Students

You also should be able to state and discuss briefly four limitations on grading and describe the ways in which your program attempts to eliminate at least two of these problems.

Limitations on Grading

- Meaning of Grades Not Consistent Across Institution or Institutions
- Inadequacy of the Evidence On Which Grades Are Based
- Detrimental Side Effects of Grades, i.e., Excessive Competitiveness, Cheating, Grades Become More Important Than Learning
- Lack of Specificity of Grades in Communicating Learning Strengths and Weaknesses

In addition, you should be able to describe two approaches to determining grades.

Approaches to Grading

- Relative-Scale Grading (Norm-Referenced)
- Absolute-Scale Grading (Criterion-Referenced)

Finally, you should be able to describe a procedure and use that procedure for determining grades for students in your course.

EXTRA LEARNING ACTIVITIES

The activities in this section will provide you with an opportunity to better understand grading, particularly in your courses. If grading policies leave you little leeway to implement your own practices, the purpose of these activities will be to help you better understand and deal with grading in your program. Hopefully this learning also will pay off for you as a college teacher.

1. Purposes of Grading

a. The purpose of grading in your course may seem quite straightforward to you. Just for fun, though, write out your grading purposes, then discuss with a colleague or two his views of the purposes of grading in their courses.

b. Try to answer by yourself or with a group of your colleagues these questions: To whom will the grades I give be reported? What decisions could conceivably be made based on these grades? How do grades I give stimulate, direct, and reward students? Ask yourself whether your answers to these questions change your attitude toward grading or the way you'll go about it.

c. How might you overcome some of the potential detrimental effects of grading? Discuss this with your colleagues in a group meeting and/or with advanced students.

2. Grading Practices

a. If you would like to know more about norm-referenced and criterion-referenced approaches to grading, take a look at Handout 8 in the Appendix.

3. Grading in Your Program

a. If you have a better way of figuring term grades, share it with your colleagues.

A P P E N D I X

INSTRUCTIONAL DEVELOPMENT WAYS AND MEANS:

Testing and Grading

HANDOUT 1

TESTING AND GRADING

BIBLIOGRAPHY

Basic Books

These paperbacks are good starting points for a first encounter with objectively scored tests.

Green, J. A.; Teacher-made Tests, New York: Harper & Row, 1968.

Gronlund, N. E.; Constructing Achievement Tests, Englewood Cliffs: Prentice-Hall, 1968.

(This book deals particularly well with different levels of learning (according to Bloom's taxonomy), using many examples.)

Wood, D. A.; Test Construction: Development and Interpretation of Achievement Tests, Merrill, 1961 (out of print).

(The only shortcoming of this book is its relative lack of explicit attention to the testing of different levels of learning. Otherwise it is excellent).

Standard References

Choices between these books are almost entirely a matter of preference for style, or availability. They are all good.

Adams, G. C.; Measurement and Evaluation in Education, Psychology, and Guidance, New York: Holt, Rinehart, & Winston, 1964.

(Chapters 10 and 11. Chapter 11 includes large numbers of items testing different levels of knowledge in a variety of subjects.)

Ahmann, J. S. and Glock, M. D.; Measuring and Evaluating Educational Achievement, Boston: Allyn and Bacon, 1971.

(Chapters 3 and 4, 4 dealing with high level learning. Chapter 11 discusses marks and grading. Quite up-to-date and extensive reference lists are included.)

Chase, C. I. and Ludlow, H. G. (eds); Readings in Educational and Psychological Measurement, Boston: Houghton Mifflin, 1966.

(Units 5 and 7. The latter includes a rare discussion of how students react in examinations and how this can be taken into account in test construction.)

Gronlund, N. E. ; Measurement and Evaluation in Teaching, London: Collier-MacMillan, 1971.

(Chapters 6 to 9. A main feature is the insistence on the relations between teaching, learning, and evaluation.)

Milton, Ohmer and John W. Edgerly, The Testing and Grading of Students, Change Magazine: New Rochelle, New York, 1976.

Nunnally, J. C.; Educational Measurement and Evaluation, New York: McGraw-Hill, 1964.

(Parts of Chapter 5 (p. 103 on), chapters 6 and 7)

Payne, D. A.; The Specification and Measurement of Learning Outcomes, Waltham: Blaisdell, 1968.

(Chapters 3 and 4. Chapter 3 of this paperback has an extensive discussion and chart (pp. 38-40) on matching the type of test to the learning situation. Chapter 10 discusses the assigning of marks and the philosophy of marking.)

Remmers, H. H., Gage, N. L. and Rummel, J. F.; A Practical Introduction to Measurement and Evaluation, 2nd ed. New York: Harper & Row, 1965.

(Chapter 8. Chapter 7 includes an interesting section on the participation of students in their own evaluation, including a college-level illustration (pp. 200-205.))

Thorndike, R. D. (ed); Educational Measurement, 2nd ed. Washington: American Council on Education, 1971.

(Chapters 2 to 4. Chapter 9 discusses the assessment of performance and products.)

HANDOUT 2

INSTRUCTIONAL OBJECTIVES:

References

1. Bloom, Benjamin (ed); Taxonomy of Educational Objectives, David McKay Company, New York 1967.
2. Duchastel, Philippe and Paul F. Merrill; "The Effects of Behavioral Objectives on Learning: A Review of Empirical Studies." Review of Educational Research, Vol. 43, #1, 1973.
3. Geis, George L; Behavioral Objectives: A Selected Bibliography and Brief Review, ERIC Clearinghouse on Media and Technology. Stanford University, Stanford, California 94305, 1972.
4. Mager, Robert F; Preparing Instructional Objectives, Fearon Publishers, Belmont, California, 1968.

TEST STATISTICS

DISTRIBUTION OF SCORES

The two most useful measures of central tendency are the mean and median. The mean is the arithmetic average. The median is the score above which and below which half of the students score. The mean and median should be approximately equal unless the distribution of scores is skewed. The magnitude of the difference between the mean and median is, therefore, an indication of the degree to which the distribution of scores is skewed.

In addition to measures of central tendency, the instructor needs some indication of how the scores vary about the mean. Two commonly used measures of variability are the range and standard deviation.

The range is simply the difference between the highest and lowest test score plus one. If the top student in a class made a score of 96 and the lowest a score of 56, the range would be 41. The range is a crude measure of variability because it does not tell you how the scores are grouped around the mean.

A better measure of variability is the standard deviation, S.D. The formula for calculation is:

$$S.D. = \sqrt{\frac{\sum(x - \bar{x})^2}{N - 1}}$$

where x = raw score, $\bar{x}$ = mean, and N = number of scores. A glance at the formula will enable the reader to notice that where $x - \bar{x}$ is large (extreme scores), the square of these values will have a significant effect on the magnitude of the standard deviation. This helps to explain why the standard deviation increases with variability in scores.

One of the most useful qualities of the standard deviation is its relationship to the distribution of scores (i.e., approximately 68% of the scores fall between plus and minus one S.D. from the mean; 95% fall between plus and minus two S.Ds. from the mean; essentially all scores lie within plus and minus three S.Ds.)

To illustrate the effect of variation on test analysis, let us examine a couple of hypothetical test results:

Test I ; mean = 60; S.D. = 10

Test II; mean = 60; S.D. = 20

Suppose a student scored 80 on both tests. Are these scores

TYPES OF TEST ITEMS

Essay, Short Answer, and Problems

Examples:

1. Outline a laboratory procedure for determining the percentage of water of hydration in hydrated copper sulfate.
2. Contrast the phlogiston and present day theories of combustion by listing 5 basic differences.
3. Find the number of grams of oxygen gas generated by the thermal decomposition of 30.0g of potassium chlorate.

Strengths:

1. Items easy to prepare.
2. Items especially effective at measuring higher order mental processes (i.e., application, analysis, synthesis, evaluation).
3. Items promote a broader, more global, approach to the study of a topic. Students who have studied for an essay test tend to do better on all types of tests than do students who have prepared for multiple choice or true-false examinations.

Weaknesses:

1. Items are usually very time-consuming and usually cannot be used to sample a broad range of content.
2. Items are difficult and time consuming to score.
3. Items are not easy to score reliably. That is, since it is difficult to be objective when scoring essay examinations, scoring of responses is somewhat subjective.

Hints on Construction and Use:

1. An essay or short answer item should present a problem to be solved.
2. The problem should be defined explicitly. This is essential to prevent ambiguity about what is expected and to facilitate the scoring process. It also inhibits "wandering", broad, unsupported generalizations.
3. The directions for the test should be very explicit. Students need to know exactly what is expected of them. For example, they need to know how much time to spend on each question, the form in which the response is to be made, and the type of information to be included in the

response.

4. A detailed key should be prepared for each question. The key is what the instructor considers to be the best answer to the item. This procedure will vastly increase the reliability (consistency) of the test item.
5. In scoring essay examinations the instructor should score the first question on each page before going to the next question. Scoring may be accomplished by sorting papers into "piles", rereading the marginal papers, and resorting. Another method often used is the Point-Score method. A key is prepared containing the important concepts and/or processes. A point value is assigned to each prior to any scoring. Papers should not be evaluated on extraneous factors such as handwriting, spelling and grammar.

Multiple Choice Tests

Examples:

1. Which of the following elements does not belong with the others?
 - a. iron
 - b. sulfur
 - c. potassium
 - d. barium
 - e. mercury
2. The addition of manganese dioxide to potassium chlorate with the application of heat will result in:
 - a. less oxygen per unit of time than when potassium chlorate is used alone.
 - b. more oxygen per unit of time than when potassium chlorate is used alone.
 - c. no change in the rate of production of oxygen.
 - d. the production of manganese chloride.
 - e. the production of potassium dioxide.

Strengths:

1. Versatility and flexibility. This type of test item may be designed to measure performance objectives at all levels of the cognitive domain.
2. The multiple-choice items may be responded to quickly, thus enabling the instructor to sample a broad range of content and objectives.
3. Items are easily and reliably scored by individuals who may not be qualified to teach the subject.
4. Patterns of wrong answers may be used to diagnose learning difficulties.

Weaknesses:

1. Items are difficult and time consuming to prepare.
2. Items cannot be used to measure the students' ability to organize materials or clearly express their answers.

Hints on Construction and Use:

1. The multiple-choice item is composed of a stem, foils or distractors, and a keyed response as in the following example:

Water is found: (--stem)

- a. in the air (--foil or distractor)
 - b. on the earth's surface (--foil or distractor)
 - c. underground (--foil or distractor)
 - d. all of the above (--keyed response)
2. Prepare the original draft items and have someone else read them to look for ambiguity and clues to the correct answer. Avoid grammatical clues to the wrong answers--singular subjects and plural verbs, etc. Items should be concise and presented in a clear and simple writing style.
 3. Items should be keyed to the objectives of the unit being tested.

True-False Items

Examples:

1. The volume of a mass of gas
 - a. tends to increase as the temperature increases (T F)
 - b. tends to increase as the pressure increases (T F)
 - c. may be held constant by increasing pressure and decreasing temperature (T F)
 - d. may be reduced by increasing pressure and decreasing temperature (T F)

Strengths:

1. Items may be rapidly generated.
2. Items are rapidly and accurately scored.
3. Items are effective for measuring the acquisition of factual information when this is desired.

Weaknesses:

1. Items are very susceptible to guessing.
2. It is difficult to diagnose learning problems from wrong responses.
3. Items are usually poor discriminators.
4. Because of the guessing factor and inability to discriminate, the true-false test is less reliable than multiple choice.
5. They tend to measure primarily objectives at the knowledge level.

Hints on Construction and Use:

1. Instructions on how to mark responses to items should be clear and unambiguous.
2. True-false tests are often modified to require the students to correct the false statements.
3. Patterns of correct responses should be randomized.
4. Statements should be positive, and words like "some" and "many" should be avoided in order to reduce ambiguity. Words like "none" and "all" often provide clues to the correct answer.

Completion Items

The completion item is a written statement which requires the student to supply the correct word or phrase in response to an incomplete sentence or question.

Examples:

1. Write the structural formula for acetic acid.
2. Aqueous barium chloride reacts with aqueous sodium sulfate to form _____ and _____.

Strengths:

1. Items are easy to construct.
2. Guessing is minimized.
3. Items are easily scored.
4. Items measure recall of information rather than recognition. When the recall of information is considered important this type of item is superior to multiple choice or matching.

Weaknesses:

1. They are restricted to use with low level objectives--primarily memory. It is extremely difficult to construct completion items that test at the analysis, synthesis, or evaluation level.
2. It is difficult to construct items that do not contain clues from the context of the statement.
3. It is difficult to construct items that have only one correct response.

Hints on Construction and Use:

1. The item should be concise, unambiguous and grammatically correct.
2. The statement should provide only one frame of reference for responding. However, if too many clues are given in an attempt to establish a single, correct response, the item may become too easy. If not enough clues are provided, the item may be too ambiguous and several responses could be considered correct.
3. Incomplete sentences should be used sparingly because they tend to be obvious or ambiguous.
4. It is generally helpful in scoring to place all blanks at the same position on the paper (i.e., on the far left or far right of the paper).

Matching Tests

The matching test is a specialized form of the multiple choice type in which a set of related stems share the same set of alternatives.

Example:

At the left you will find a list of chemical names of common chemical compounds. At the right you will find a list of common names for these compounds. Place the letter of the common name in the blank in front of the corresponding chemical name.

- | | |
|-----------------------------------|----------------|
| ___ 1. methyl salicylate | a. quicksilver |
| ___ 2. sodium bicarbonate | b. epsom salt |
| ___ 3. ferric oxide | c. gasoline |
| ___ 4. mercury | d. baking soda |
| ___ 5. acetic acid | e. rust |
| ___ 6. octane | f. vinegar |
| ___ 7. hydrated magnesium sulfate | g. aspirin |
| ___ 8. acetylsalicylic acid | h. wintergreen |
| | i. limewater |
| | j. sugar |

Strengths:

1. Items are quite good for measuring at the lower levels of the cognitive domain. Factual-memory level material like vocabulary, formulas and single relationships can be measured effectively with this kind of item.
2. Items are scored easily and rapidly.

Weaknesses:

1. It is difficult to construct a matching test that measures at the higher levels of the cognitive domain.
2. Since the items must have a degree of homogeneity, there is a danger of concentrating a disproportionate amount of attention on selected content areas.

Hints on Construction and Use:

1. Matching items are arranged in groups of related statements and a corresponding list of alternatives.
2. An introductory statement must be included which describes the relationship between the stems and the corresponding alternatives, and which indicates how the student should respond.
3. To reduce the probability of students obtaining correct answers by guessing, it is preferable to include more alternatives than stems.
4. The stems should exhibit homogeneity in their relationship to each answer. Unrelated stems will provide clues to the correct answer.
5. Make sure that the items relate to the unit objectives.
6. Each of the alternatives must be grammatically appropriate to each stem in the set.
7. Sometimes it is appropriate to note in the instructions that an alternative may be used more than once.
8. The lists of stems and alternatives should not be too long. Approximately ten items should be considered as a maximum for a matching test.
9. Be careful to avoid a pattern in the order of stems and keyed responses.

Performance Tests

Performance tests are procedures for evaluating a student's ability to identify objects, manipulate materials or equipment, perform assigned tasks, or react to real or simulated situations. In chemistry this might involve the analysis of an unknown, the generation of a chemical product, titration to a given degree of accuracy, etc.

Unfortunately, most classroom testing is accomplished in written form. There are many important instructional outcomes that are difficult or impossible to measure in this manner. Skill in the manipulation of expensive laboratory apparatus like an analytical balance, a pH meter, or an infrared spectrophotometer is an example.

Written tests tend to focus on knowledge and understanding of concepts. A performance test usually measures the ability to apply concepts and principles to new and changing conditions.

Example:

On the lab benches you will find a standard solution of 0.1003M KOH and an HCl solution of unknown concentration. Using your best lab technique, perform a titration and calculate the concentration of the HCl.

Strengths:

1. They can be used to measure objectives where written tests are inappropriate.
2. They can be used with students who lack the verbal ability to cope with written exams.

Weaknesses:

1. They often require a large amount of time to prepare and administer.
2. They are usually difficult to score objectively.
3. It is often difficult to obtain an adequate sample of behavior.
4. In some situations only one student can be examined at a time.

Hints on Construction and Use:

1. The examiner must be careful to select only the most important and critical behaviors to measure.

2. The conditions under which the test is administered should be carefully controlled.
3. The scoring procedure should be as objective as possible. However, in measuring performance, it is often difficult to divorce subjective judgment from objective measurement. For example, if lab technique is being measured some judgment of quality is necessary.
4. Subjective judgments may be made more objective by a careful delineation of the characteristics of the desired behavior--often in the form of a rating scale.
5. It is helpful to break down a complex behavior (e.g., good titration technique) into small steps. A relative value (e.g., points) may be assigned to each step in proportion to their value in contributing to the overall desired behavior.
6. The examiner must determine what is to be scored--process, product, or both. If process is to be evaluated, the examiner must decide if quality and accuracy are as important as speed. If all are considered important, appropriate weights must be assigned to each.

HANDOUT 5

TABLE OF SPECIFICATIONS

From: Mehrens, William A. and Lehmann, Irvin J., Measurement and Evaluation in Education and Psychology, Holt, Rinehart and Winston, Inc. 1973.

Table of Specifications

One of the major complaints made by students with regard to teacher-made tests is that they are often invalid. Students may not use the technical

* The table of specifications is sometimes called the test blueprint, test grid, or content validity chart.

term "validity," but their comments—"we were tested on minute, mundane facts," "the material we were tested on wasn't covered in class nor is it in the text or assigned readings," and "we spent one-half of the course discussing the concepts of validity and reliability and yet only 5 percent of the test was devoted to these concepts"—all point out that the test lacked content validity. We recognize that there may be some instances where students are only displaying a "sour grapes" attitude and complain for the sake of complaining. But, we know from looking at a plethora of teacher-made achievement tests that there is some justification for the complaints made. Although a table of specifications is no guarantee that the errors will be corrected, such a blueprint may help improve the content validity of teacher-made tests.

Did you ever see an engineer build a structure without using a blueprint? Did you ever see a plumber or electrician work in a new home without referring to some type of blueprint? Did you ever see a tailor make a suit without a pattern? We haven't. But we have seen some tests constructed by teachers and college professors who did not use some type of plan or guide or blueprint, and their tests reflected it.

The purpose of the table of specifications is to define as clearly as possible the scope and emphasis of the test, and to relate the objectives to the content. For example, how does one proceed to construct a test on American history or literature or mathematics? The area to be tested must first be delimited further. The test must be balanced and should reflect the objectives of instruction. In other words, the test must be a valid measure of the pupil's knowledge, skills, and other attributes that the teacher sought specifically to teach. By ensuring that the test adequately covers the content and is directly related to the objectives, the teacher has satisfied the most important criterion of the test—content validity.

Who Should Prepare Specifications?

Some teachers argue strongly against involving students in the development of the table of specifications. These teachers contend that if students play a role in the decision-making process, education will be eventually run by the students rather than for them. We see nothing wrong in involving students in the development of the table of specifications. In fact, where feasible, we encourage such student involvement, if for no other reason than to have students feel that they have played some role in planning the course. Our attitudes toward student participation should *not* be interpreted as condoning students' complete control. Nor should our remarks be regarded as an abrogation of the teacher's major and final responsibility. The *teacher* is the decision maker—not the students. Student input should be considered by the teacher in making his decisions and should be utilized only in an *advisory* capacity.

When to Prepare Specifications

Should the table of specifications be prepared just before the test is constructed or should it be prepared before the course has been taught? Ideally,

and to be of most benefit, the table of specifications should be prepared *before* beginning instruction. It would be nice if it were considered part and parcel of the course preparation. Why? Because these "specs" may help the teacher be a more effective teacher. These "specs" should assist the teacher in organizing his teaching material, his outside readings, his laboratory experiences (if necessary)—all the resources he plans on using in teaching the course. In this way, the "specs" can help provide for optimal learning on the part of the pupils and optimal teaching efficiency on the part of the instructor. Hence, if the "specs" are to assist and be of value to the classroom teacher, they should be prepared *before* the instructional phase is begun. In a way, then, the "specs" serve as a monitoring agent and can help keep the teacher from "straying off" his instructional track.

Preparing the Table of Specifications

Once the course content and instructional objectives have been specified, the teacher is ready to integrate them in some meaningful fashion so that the test, when completed, will be a valid measure of the student's knowledge. Table 7-3 is an extension of the one-way grid of the course content in Natural Science (Table 7-2) and incorporates the course objectives. This extension is a two-way grid that simultaneously relates the course content to the behavioral objectives as defined by Bloom et al. (1956).

One could, of course, delineate the course content into finer subdivisions. Whether this needs to be done depends upon the nature of the content and the manner in which the course content has been outlined and taught by the teacher. A good rule-of-thumb to follow in determining how detailed the content area should be is to *have a sufficient number of subdivisions so as to ensure adequate and detailed coverage*. The more detailed the blueprint, the easier it is to get ideas for test items.

You will notice in Table 7-3 that there are numbers in certain cells and blanks in other cells. If you were to total the last column or bottom row you would get 100. Now, what do all these numbers mean? The number 100 in the bottom right-hand cell (corner) is the total percentage (it can be, however, related to the number of items on the test) or numerical value or worth of the test. The numbers at the bottom of each column indicate the percentage of the test devoted to a particular objective. Hence, in this hypothetical test, 25 percent of the test items measured knowledge, 30 percent of the items measured comprehension, and so forth. The numbers in the last column signify the percentage of items that were allocated to each content area. In this case it was decided that each content area would be weighted equally, and hence 10 percent of the items were written for each content area. The boldface number 5 tells you that 5 percent of the total test was devoted to the measurement of "knowledge" in methods of science. It also tells you that, for methods of science, 50 percent of the items written in this content area measured the knowledge objective. At this point you might ask: "Who

TABLE 7-3 Two-Way Table of Specifications for a Final Examination in Natural Science

Course Content	Objectives*				TOTAL
	KNOWLEDGE	COMPREHENSION (Translation Interpretation, Extrapolation)	APPLICATION	ANALYSIS	
1. Methods of science; hypotheses concerning the origin of the solar system	5	2		3	10
2. Minerals and rocks	5	5			10
3. Changes in land features	4	4	2		10
4. Interpretation of land features	2	2	6		10
5. Animal classifications	2	4	4		10
6. Plants of the earth	4	4	2		10
7. Populations and the mechanisms of evolution	3	3		4	10
8. Variation and selection		1	5	4	10
9. Facts of evolution and the theory that explains them		2	2	6	10
10. Evolution, genetics, and the races of man		3	4	3	10
Total	25	30	25	20	100

SOURCE: Nelson, 1958.

* Objectives are based on Bloom's taxonomy.

determines the weights?"—that is, who determines the proportion of items that are designed for each content area and for each objective?

Determination of Weights

You will recall that one of the major advantages of the teacher-made versus commercially published test is that the teacher-made test can be "tailor-made" to fit the teacher's unique and/or particular objectives. In this way, Mr. Molecule, who wishes to stress the gas laws in eleventh-grade physics, can do so while Mr. Element, who wishes to stress heat and mechanics, is also at liberty to do so. Each teacher can prepare a test that is valid for his students. Because the classroom teacher—more so than any other person—knows the relative emphasis placed upon the various instructional objectives,

it naturally follows that he should have the major responsibility in assigning the various weights to the cells in Table 7-3. There is no hard-and-fast rule that can be prescribed for the teacher to use in determining the weights to be assigned to the various cells in the table of specifications. Experience is his best resource. All that we can say is that the weights assigned should reflect the relative emphasis used by the teacher when he taught the course.

As an example we give a very simplified illustration of how a classroom teacher can initially determine the weights to be assigned to a particular cell in the table of specifications. Assume that Mr. Atom will spend five class periods on a unit in hydrogen and he wants to prepare a test on this unit. Mr. Atom plans to spend one period (20 percent of the time) discussing the physical properties of hydrogen; one and one-half periods (30 percent) on the chemical properties of hydrogen; one-half period (10 percent) on the preparation of hydrogen; and two periods (40 percent) discussing the uses of hydrogen. These values are represented as the row totals in Table 7-4. In teaching this unit, Mr. Atom will be concerned with three behavioral changes: the ability to recall information; understanding of basic concepts and principles; and the ability to apply information, concepts, and principles in new situations. The relative emphasis placed on each of these behavioral (instructional) objectives will be 40, 30, and 30 percent, respectively. These values are represented as the column totals. Mr. Atom must now assign values to each of the 12 cells. This is done by multiplying the row totals by the column totals. For example, the cell involving recall of information in physical properties would have a weight of $.20 \times 40 = 8\%$; the cell incorporating application and uses would have a weight of $.40 \times 30 = 12\%$. This procedure is repeated for each cell and is illustrated in Table 7-4. Mr. Atom now has a blueprint *both* to guide him in teaching this unit and in constructing a test on this unit.

There arises now the question of how firm the assignment of weights in each cell should be. We take the position that the initial weights in each cell should be considered as tentative. It is only after the course has been taught

TABLE 7-4 Table of Specifications for a Chemistry Unit on Hydrogen

Content, %	Instructional (Behavioral) Objective			TOTAL, %
	RECALL OF INFOR- MATION	UNDER- STANDING CONCEPTS	APPLICA- TION IN NEW SITUATIONS	
Physical properties	8	6	6	20
Chemical properties	12	9	9	30
Preparation	4	3	3	10
Uses	16	12	12	40
Total	40	30	30	100

that the weights can be considered as definite. And, because conditions may vary from one class to another, the final weights may be somewhat different for different classes taught by the same teacher. This, however, does not mean that the teacher should depart from his original "specs" because he finds it difficult to write items designed to measure the higher mental processes. As Tinkelman (1971, p. 56) wrote: "If a test blueprint rests upon a sound judgmental basis, the test constructor has the professional obligation to obtain items of satisfactory quality and in sufficient numbers to satisfy blueprint specifications."

There are some instances where following the original table of specifications slavishly could conceivably result in an invalid test. For example, the teacher of eleventh-grade American history might find that about 25 percent of the course was spent on the Civil War, even though he originally planned to spend only 10 percent of the time on this unit. Does this mean that he should only have 10 percent of his examination devoted to questions dealing with the Civil War? We think not. It is for reasons such as these that we feel the initial weights assigned should be considered as being flexible. But, again, flexibility should not be interpreted as license to arbitrarily change the weights without a good reason. And, if changes are made, students should be told.

Without a well-thought-out and prepared table of specifications, there is a great possibility that the test, when finally constructed, will lack content validity. If you look at the typical teacher-made final examination constructed by, say, a social studies teacher, you would see that many items test for the recall of names, places, and dates. Does this mean that the social studies teacher is not concerned with developing critical thinking? Does this in fact imply that the teacher stressed *only* the knowledge of isolated facts and skills? Not necessarily! The apparent undue emphasis on recall of factual material or computational skills that occurs in so many teacher-made tests probably occurs because of the difficulty of writing test items that measure the higher mental processes of comprehension, application, analysis, and interpretation.

Tables of Specifications for Criterion-Referenced Tests

As we mentioned in Chapter 3, the major difference between criterion- and norm-referenced tests is whether we interpret a person's score by comparing it with a specified behavioral criterion of proficiency (90 percent of the items answered correctly) or by comparing it with the scores of other people (Allan did better than Jack). However, *both* norm- and criterion-referenced test scores are related to *content*; therefore, in building an achievement test whose scores will be norm-referenced, we must be as concerned with content validity as when building a test whose scores will be criterion-referenced. In both cases, there is a domain of relevant tasks or behaviors from which we wish to sample. In both cases, we should use a table of specifications to ensure that our sample of test items is representative of the domain of behaviors. The major difference is that for the specific types of instructional decisions where

one usually finds criterion referencing to be of more value, the content domain is quite narrow or limited in focus. For those educational decisions where one is likely to find a normative interpretation of the score most useful, the domain of tasks is usually more broad. Thus, when building a test whose scores will be interpreted by comparison with some specified criterion, we may well have only one cell in the table of specifications. The domain is so narrow that it need not be subdivided. For example, we might build a criterion-referenced test on the task of adding two single-digit whole numbers. A table of specifications for such a test would not need to be subdivided. If a test had a broader focus—that is, was designed to measure addition of whole numbers, fractions, and decimals—then one would build a table of specifications so that each subtype of addition problem was represented. This would be true whether one wished to interpret the scores in an absolute or relative fashion.

In summary, the use of a test blueprint or table of specifications will help ensure that (1) only those objectives actually involved in the instructional process will be assessed, (2) each objective will receive a proportional emphasis on the test in relation to the emphasis placed on that objective by the teacher, and (3) no important objective or content area will be inadvertently omitted. Much time and effort are (or can be) expended in preparing a table of specifications. In the long run, however, it will be time and effort well spent, as it can aid immensely in the preparation of test items, in the production of a valid and well-balanced test, in the clarification of objectives to both teacher and students, and in assisting the teacher to select the most appropriate teaching strategy. Better classroom tests would be built if more careful attention were paid to the precise statements of test specifications (Ebel and Damin, 1960, p. 1508).

Distributing the Table of Specifications

Naturally, the table of specifications should be used by the teacher. But there is a difference of opinion as to whether it should be given to the students. We contend that *it should be given to the students* (especially those in the upper elementary grade and above) at the first class meeting, and should be discussed thoroughly with them.⁹ This can help minimize, if not eliminate, future misconceptions, misunderstandings, and problems.

Too frequently teachers prepare or outline their course of instruction and fail to distribute it to the students. Why? Are they ashamed of it? If they are, it cannot be very good and should be revised immediately. And, if it *is* good, is the student entitled to know his teacher's objectives? Definitely yes! Is it not conceivable that a prior knowledge of the table of specifications might result in the student's placing less emphasis in those areas that would contribute

⁹ This would also allow the students to voice opinions concerning course content and the relative emphasis. If changes are made as a result of this interaction, the pupils should be given a revised set of "specs."

little to his total score, thereby concentrating on the more important objectives of instruction? Yes, he might! Is this bad? Not necessarily so. If it will help motivate the student to concentrate on the major course objectives isn't this good? We think so.

Using the Table of Specifications

We have already discussed how the table of specifications can assist the teacher vis-à-vis his instructional strategy and in preparing valid evaluation instruments. Especially since the "age of accountability" there has been some concern voiced by educators that teachers will be prone to "teach for the test." It could therefore be interpreted that teachers who prepare and follow a table of specifications may be prone to follow this practice. As we said in Chapter 2, if the teacher has an appropriate set of instructional objectives, he should teach his pupils to realize these objectives. It is not wrong to "teach for the test" in this sense. In fact, we would be most pleased if teachers would take the time to develop appropriate instructional objectives, teach them, and then test to see the extent to which they were realized by their pupils. This is markedly different from teaching the test items per se.

THREE-DIMENSIONAL GRID FOR TEST ITEM CATEGORIZATION

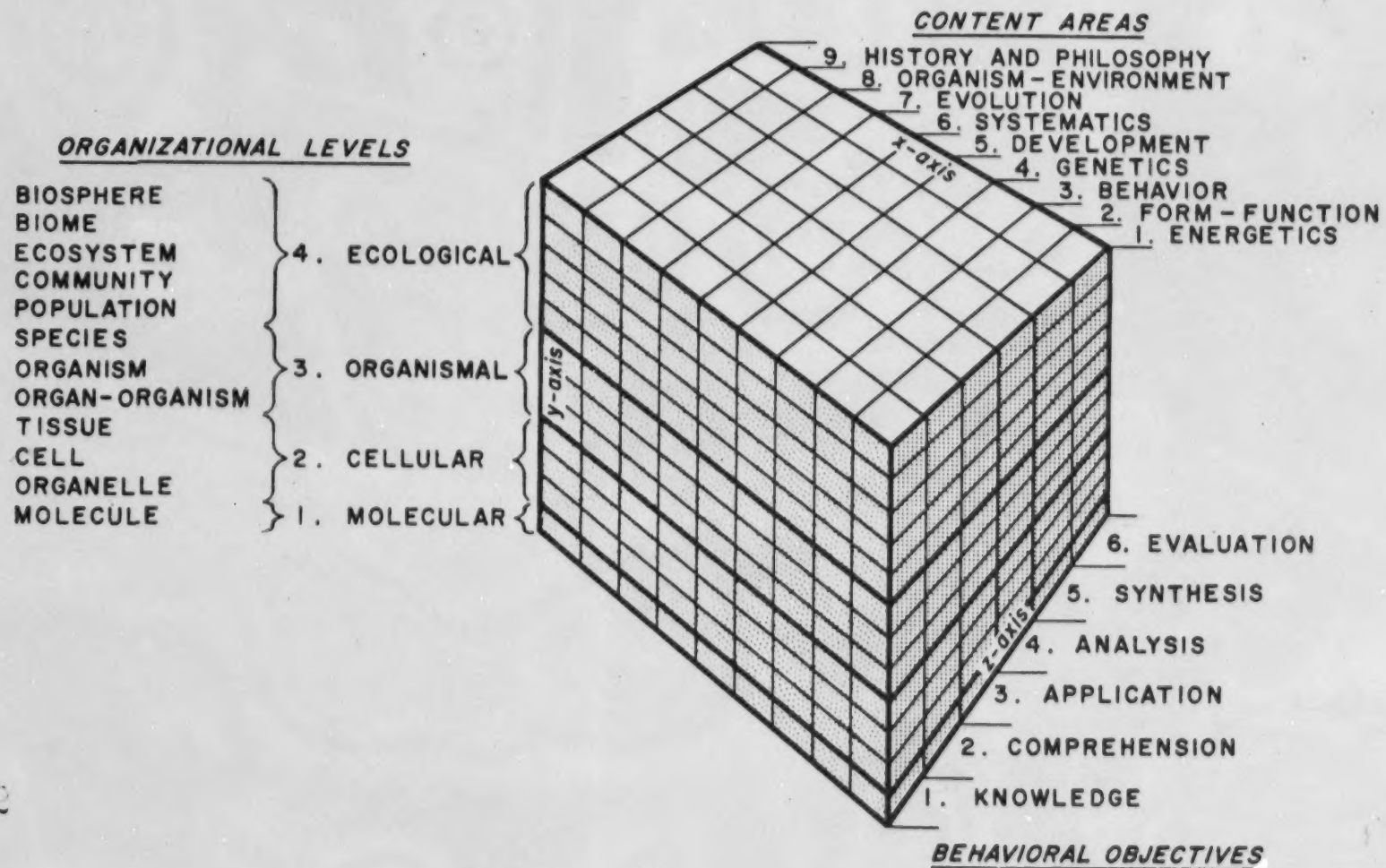


Figure 1. Three-dimensional grid to serve as a guide in the creation and categorization of test exercises.

HANDOUT 6

RELIABILITY AND VALIDITY

From: Eble, Robert L., Essentials of Educational Measurement,
Prentice-Hall, Inc., Englewood Cliffs, New Jersey 1972.
pp. 407-410, 435, 447.

A very convenient conception is that of the "reliability coefficient"
of any system of measurements for any character. By this is
meant the coefficient between one half and the other half of
several measurements of the same thing.

C. SPEARMAN

15

HOW TO ESTIMATE, INTERPRET, AND IMPROVE TEST RELIABILITY

THE IMPORTANCE OF RELIABILITY

For most tests of educational achievement, the reliability coefficient provides the most revealing statistical index of quality that is ordinarily available. If the scores yielded by any educational achievement test were all perfectly accurate, true scores with no errors attributable to the particular sample of questions used, to alertness, anxiety, fatigue, or other factors that might affect examinee performances, to lucky guesses or unlucky slips, and with no errors caused by the mistakes or biases of the person scoring the test, then the test would have perfect reliability, reflected by a reliability coefficient of 1.00. No educational achievement test, no other type of mental test, and indeed no physical measurement has ever achieved this degree of perfection. Error is unavoidably involved in any measure-

ment, but the goal of measurement specialists in all fields is to reduce these inevitable errors of measurement to a reasonable minimum.

Expertly constructed educational achievement tests often yield reliability coefficients of 0.90 or higher. In contrast, the achievement tests used in many elementary, secondary, and college classrooms often show reliability coefficients of 0.50 or lower. One of the ways of making test scores more reliable is to lengthen the test on which they are based, that is, to include more questions or items in it and to allow more time for it. But a test having a reliability coefficient of 0.50 would need to be increased to nine times its original length to bring its reliability up to 0.90. (The basis for this statement is provided by formula 15.1 on page 413.) Hence, from this point of view, a test having a reliability coefficient of 0.90 is nine times as good as a test having a reliability coefficient of 0.50.

Reliability is a necessary but not sufficient condition for quality in an educational achievement test. The author of a test that yields highly reliable scores may only have succeeded in measuring something irrelevant or trivial with very great precision. On the other hand, the author of a test that yields only unreliable scores has clearly not succeeded in measuring anything very precisely. If a test does not yield reliable scores, whatever other potential merits it may have are blurred and may be largely lost. Only to the degree that test scores are reliable can they be useful for any purpose whatsoever.¹

Reliability is important to students whose grades are often heavily dependent on the scores they make on educational achievement tests. If they were clearly aware of the importance of test reliability to them, it is likely that they would ask for evidence that the tests used to measure their achievement are not only fair in terms of the purposes of the course, but also are of sufficient technical quality to yield reliable scores.

Reliability is important also to the teacher who is aware that his examinations have shortcomings and who seeks to improve them. Estimates of the reliabilities of his tests would provide the essential information for judging their technical quality and for motivating efforts to improve them. Lengthening an unreliable test is not the only way, and may not be the best way, to improve its reliability. If modern knowledge and techniques of test construction are applied, most educational achievement tests can be made to yield scores having reliability coefficients that at least approach 0.90.

¹Robert L. Ebel, "The Value of Internal Consistency in Classroom Examinations," *Journal of Educational Measurement*, vol. 5 (1968), 71-73.

THE CONCEPT OF TEST RELIABILITY

One way of clarifying the concept of reliability is to contrast it with validity. These two definitions point up the contrast.

1. The term "reliability" means the consistency with which a set of test scores measure whatever they do measure.
2. The term "validity" means the accuracy with which a set of test scores measure what they ought to measure.

If the perforations on a target made by successive shots from a rifle are all clustered closely, the rifle is performing reliably. If those perforations are all clustered in the bull's-eye, the rifle is also performing validly.

The number of words in a poem can be measured with high reliability. Such a measure would be readily accepted as a valid measure of the length of the poem. It would not, however, be accepted by most poets or literary critics as a valid measure of the literary merit of the poem.

For most tests of educational achievement no good statistical measure of validity is ordinarily obtainable. To estimate its validity, we must look at the test itself, and, if they are available, at the principles used by the test constructor in deciding what kind of tasks to include in the test. On the other hand, a good estimate of the reliability of the test can be obtained quite simply. Since a test must be reliable if it is to be valid, the statistical analysis of the quality of educational achievement test scores ordinarily places primary emphasis on the reliability of those scores. A test yielding reliable scores is not necessarily valid. But an educational achievement test composed of questions that seem pertinent and significant to expert teachers or professors is very likely to be as valid as it is reliable. This is especially likely if those instructors also have some expertness in educational measurement.

At least three factors contribute to the reliability or unreliability of a set of test scores. One is the appropriateness and definiteness of the task. Tasks that are too easy or too difficult or permit students to make widely divergent interpretation of what is expected of them are not likely to yield highly reliable scores. Another is the constancy or stability of a student's ability to perform the tasks presented in the test. Human beings vary from hour to hour and from day to day in their alertness, energy, emotional balance, and other characteristics. If these personal variables affect a student's test performance appreciably, they will reduce the reliability of the scores. A third is the consistency and objectivity of the person who scores the test. To the extent that the scores he assigns depend upon his notions of the moment rather than on consistent standards,

uniformly applied to all the papers he reads, the scores he records will lack reliability.

AN OPERATIONAL DEFINITION OF TEST RELIABILITY

The reliability coefficient for a set of scores from a group of examinees is the coefficient of correlation between that set of scores and another set of scores on an equivalent test obtained independently from the members of the same group.

Three aspects of this definition deserve comment. First, it implies that reliability is not a property of a test by itself but rather of a test when applied to a particular group of examinees. The more appropriate a test is to the level of abilities in the group, the higher the reliability of the scores it will yield. The wider the range of talent in a group, the higher the reliability of the scores yielded by a test of that talent.

Second, the operational definition specifies the use of a correlation coefficient as a measure of reliability. One of the properties of the correlation coefficient is that it provides a relative, rather than an absolute, measure of agreement between the pairs of scores for the same persons. If the differences between scores for the same person are small relative to the differences between scores for different persons, then the test will tend to show a high reliability. But if the differences between scores for the same person are large relative to the differences between persons, then the scores will show low reliability.

Third, the operational definition calls for two or more independent measures, obtained from equivalent tests of the same trait for each member of the group. This is the heart of the definition. The different means by which sets of independent measurements of the same achievement are obtained provide the differences between the various methods for estimating test reliability.

Examinations comprise a crucial aspect of the whole educational process, since they represent willy-nilly what all the fine words have been about. Moreover examinations furnish practically the only objective evidence of the value of a course of instruction. A published examination (and examinations are always published, whether officially or unofficially) presents a fair statement of the course objectives. All else that is said about the course may be summarily classified in the category of "pious hopes."

M. W. RICHARDSON

16

THE VALIDITY OF CLASSROOM TESTS

THE CONCEPT OF VALIDITY

In the preceeding chapter the concept of validity was distinguished from that of reliability in these two parallel but different definitions.

1. The term "validity" means the accuracy with which a set of test scores measures what it ought to measure.
2. The term "reliability" means the consistency with which a set of test scores measures whatever it does measure.

Having defined reliability in Chapter 15, we proceeded to show how reliability coefficients could be calculated and interpreted. It would be nice if we could treat validity in the same way. Unfortunately, for most tests that are of interest to classroom teachers, quantitative measures of test validity cannot be obtained. Only when tests are used to predict subse-

quent performance, or used as simpler substitutes for some exact but difficult measurement process, can coefficients of validity ordinarily be obtained. The validity of most classroom tests is a matter of carefully defining what is to be measured, and of relying on expert judgment to determine how well the tests measure it.

Solid evidence supporting the validity of most educational tests is very difficult to obtain. Possibly because of this, test specialists have worried a great deal about the validity problem and have written many articles about it. In contrast, the physical scientists, who probably do a better job of measuring than do psychologists and educators, seem to be relatively indifferent to, even unaware of, the validity problem. They tend to worry more about how accurate and how useful their measurements are than about essentially unanswerable questions which ask whether they are measuring what they *intend* to measure or what their procedures *purport* to measure.

Consider this question, "Which is the more valid measure of central tendency, the mean or the median?" Surely this question has much in common with the question, "Which is the more valid measure of intelligence, the Stanford-Binet or the Wechsler scale for children?" Why is it that the first question is almost never asked, whereas the second is asked frequently? Neither question, it may be pointed out, has ever been answered conclusively. In neither case do we have any independent external criterion to which we can appeal for a decision.

With respect to the first question, our behavior is more like that of the typical scientist; that is, we define the measurements, develop their properties, and use whichever is more appropriate or convenient in a specific situation. With respect to the second, we withhold full confidence in either test in the hope that someone will find an answer to the essentially unanswerable question we have asked. Neither central tendency nor intelligence is a meaningful quantitative concept apart from some specific operations for calculation or some specific testing procedures.

TWO TYPES OF VALIDITY

A great many varieties of validity have been discussed in the rather extensive literature on this subject. Here are brief characterizations of several of these varieties.

"*Concurrent validity* is concerned with the relation of test scores to an accepted contemporary criterion of performance on the variable that the test is intended to measure."¹

¹American Educational Research Association, *Technical Recommendations for Achievement Tests* (Washington, D.C.: AERA, 1955), p. 16.

Construct validity is concerned with "what psychological qualities a test measures," and one evaluates it "by demonstrating that certain explanatory constructs account to some degree for performance on the test."¹

"Content validity is concerned with the adequacy of sampling of a specified universe of content."²

"Curricular validity is determined by examining the content of the test itself and judging the degree to which it is a true measure of the important objectives of the course, or a truly representative sampling of the essential materials of instruction."³

Empirical validity "refers to the relation between test scores and a criterion, the latter being an independent and direct measure of that which the test is designed to predict."⁴

"Face validity refers, not to what a test necessarily measures, but to what it appears to measure."⁵

"The factorial validity of a test is the correlation between that test and the factor common to a group of tests or other measures of behavior . . . such validity is based on factor analysis."⁶

Intrinsic validity involves the use of experimental techniques other than correlation with a criterion to provide objective, quantitative evidence that the test is measuring what it ought to measure.⁷

"Predictive validity is concerned with the relation of test scores to measures on a criterion based on performance at some later time."⁸

"Validity by definition: For some tests the objective is defined solely in terms of the population of questions from which the sample comprising the test was drawn, for example, when the ability to handle the 100 number facts of addition is tested by sampling of those number facts."⁹

These varieties of validity are not all distinctly different from each other. In fact, one or two of them are practically identical with one or two others. But enough major differences appear to justify grouping

¹American Psychological Association, Inc., *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (Washington, D.C.: APA, 1954), p. 14.

²*Technical Recommendations*, p. 16.

³C. C. Ross and Julian C. Stanley, *Measurement in Today's Schools* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1954), p. 101.

⁴Anne Anastasi, *Psychological Testing* (New York: The Macmillan Company, 1954), p. 127.

⁵*Ibid.*, p. 121.

⁶*Ibid.*, p. 123.

⁷Harold Gulliksen, "Intrinsic Validity," *The American Psychologist*, vol. 5 (October 1950), 511.

⁸*Technical Recommendation*, p. 16.

⁹Charles I. Mosier, "A Critical Examination of the Concepts of Face Validity," *Educational and Psychological Measurement*, vol. 7 (Summer 1947), 192.

them into two major categories: those concerned with primary or direct validity and those concerned with secondary or derived validity. A test has direct, primary validity to the extent that the tasks included in it represent faithfully and in due proportion the kinds of tasks that provide an operational definition of the achievement or the trait in question. A test has derived, secondary validity to the extent that the scores it yields correlate with criterion scores that possess direct, primary validity. Thorndike and Hagen suggest a similar dichotomy of types of validity: those which depend primarily on rational analysis and professional judgment (here identified as *direct* validity) and those which depend on empirical and statistical evidence (which are designated here as *derived* validity).¹¹

Some types of validity that seem appropriate for each category are listed below:

<i>Direct</i>	<i>Derived</i>
Validity by definition	Empirical validity
Content validity	Concurrent validity
Curricular validity	Predictive validity
Intrinsic validity	Factorial validity
Face validity	Construct validity

The distinction between the two categories is not sharp in all cases. For example, it can be argued that factorial validity and construct validity, despite their involvement of multiple measurements and coefficients of correlation, do represent a basic (primary) kind of validity. Still, a distinction seems warranted between primary validity, which must be built into a test and which can be evaluated only by examining critically the decisions of the test constructor, and derived validity, which always involves correlation with some real or hypothetical criterion.

DERIVED VALIDITY

There are two kinds of derived validity that are of principal interest to constructors and users of educational tests: concurrent and predictive. Correlations between scores on a test and criterion measures available at the time the test is given indicate concurrent validity. Correlations between prior test scores and subsequent measures of achievement indicate predictive validity.

¹¹Robert L. Thorndike and Elizabeth Hagen, *Measurement and Evaluation in Psychology and Education* (New York: John Wiley & Sons, Inc., 1955), pp. 109-10.

Efforts are sometimes made to determine the concurrent validity of educational achievement tests by correlating the test scores of pupils with the grades teachers assign to the same pupils. If the test scores *were* used to determine grades the correlation would have to be high, whether the test were any good or not. Part of the trouble is that teachers' grades often are not highly reliable. Another part is that the grades usually do, and usually ought to, reflect pupil achievements that the test could not or was not intended to measure. One could reasonably expect students who get grades of A in a biology course to make a higher average score on a biology achievement test than the students who get grades of B. But it would be unreasonable to expect every A student to make a higher score on the test than any B student. Further, one could probably show that the average score of A students in biology on an intelligence test was higher than the average score of the B students. Would this then validate the intelligence test as a biology test?

Predictive validity applies more to aptitude tests than to classroom tests of educational achievement. A case can be made for the contention that all measurement is for purpose of prediction, but this does not alter the fact that there is a great difference between the way aptitude test scores and achievement test scores are typically used. The validities of classroom tests of educational achievement are seldom judged on the basis of how well they predict subsequent achievement.

OPERATIONAL DEFINITIONS

For many achievement tests the most important type of validation, sometimes the only possible type, is direct validation. Rulon has stressed the primary role of direct intrinsic validity for classroom tests in these words.

What we need is to be able to choose between available test techniques on the basis of what operations we are trying to teach the learner to perform, and what materials we are trying to teach him to perform these operations upon. Both these materials and these operations should be represented in the test situation if the test is to be "obviously valid."

Such a test must always be the criterion by which any not obviously valid test is validated.

The "obviously valid" test is its own criterion.¹²

All types of direct validation require explicitly or by implication a definition of the achievement or trait to be measured. An essential

¹²P. J. Rulon, "On the Validity of Educational Tests," *Harvard Educational Review*, vol. 16 (1946), 290-96.

characteristic of a measurable property, as Cook has pointed out, is that it must be clearly defined and unambiguous.¹³ "So long as we stick to operational definitions in discussing and thinking about the abstract concepts used in educational measurements," says Cureton, "we will be on safe ground."¹⁴ To be operationally defined, an achievement must be defined in terms of tasks that will differentiate between those having more and those having less of that achievement.

There is no better way of making clear what one means by achievement in algebra or chemistry or psychology than by describing how one would measure the amounts of those achievements that other persons possess. Much thinking and writing about educational goals and outcomes is vaguer than it ought to be because words like "intelligence," "motivation," and "creativity" are used freely without definite specification, and probably without any definite conception, of their operational definitions. Good tests of human traits can provide useful operational definitions of those traits.

But many classroom tests do not provide very clear or authoritative definitions of the qualities they measure. They reflect the understandings and values and limitations of a single test author. Even though this test constructor has carefully thought out the basic rationale for the test (and many constructors of classroom tests fail to do this, unfortunately), this rationale may not be clearly evident to one who looks at the test. Thus while one can say that every test does define the trait it measures in some fashion, most classroom tests do not define what they measure very authoritatively or clearly. Usually they would be better tests if their authors would try to make them better operational definitions of what they ought to measure.

Many of the problems of validation that perplex test constructors are really problems of defining operationally the achievement or the trait to be tested. If anxiety, for example, were defined authoritatively, comprehensively, and unequivocally, in terms of the responses to standard test situations that differentiate more anxious from less anxious persons, there would be far less concern about the validity of an anxiety test. The same statement can be made about practically all other traits or achievements that instructors would like to test. The persistent difficulties that plague our efforts to measure some outcomes of education are less attributable to limitations of measurement than to uncertainty about what is to be measured.

Well, then, why not get busy defining in operational terms these

¹³Walter W. Cook, "Tests: Achievement," in *Encyclopedia of Educational Research*, ed. Walter S. Monroe, rev. ed. (New York: The Macmillan Company, 1950), p. 1464.

¹⁴E. E. Cureton, "Validity," in *Educational Measurement*, ed. E. F. Lindquist (Washington, D.C.: American Council on Education, 1951), p. 625.

traits and achievements in which we are interested? We should, but we should not look for quick, easy success. Two obstacles, at least, may be expected. One is the persistent belief that what a word (or a concept) like "anxiety" or "intelligence" or "creativity" means can somehow be discovered by research or by the analysis of data. But language is always invented, and what a term means is always determined by some consensus among past and present users of the term. Research can discover what people *do* mean by the term "anxiety." It cannot reveal what anxiety *ought* to mean. Tests cannot reveal what intelligence *really* is; they can only illustrate what certain test constructors believe that the term "intelligence" ought to mean.

But different tests thus imply different definitions of intelligence or of almost any other trait or achievement for which tests have ever been made. This suggests the second obstacle—the difficulty of reaching substantial agreement on the operational definition of a trait or an achievement. If agreement is not reached, if each test constructor projects his own unique perceptions and values and theories in each test he constructs, without regard for the perceptions and values and theories of others who build similar tests, nothing of general value has been defined and no test of general validity is likely to be produced.

Fortunately, our inability to discover in nature what a term should mean or the absence of any revelation of ultimate criteria for our tests does not condemn us to an endless Babel of competing operational definitions. Not all possible operational definitions of an achievement are likely to be equally reasonable. They are not likely to yield tests that prove equally useful. They are not likely to be publicized with equal vigor and skill. If the concept to be defined is necessary and useful, a consensus on its meaning will emerge in time. The process can be hastened if enough of those concerned with using the concept will dedicate enough of their time and energies to developing a rational consensus. Substantial progress has already been made in defining some traits and achievements in these terms. A great deal more would be useful.

THE ROLE OF JUDGMENT

The dependence of direct intrinsic validity on human judgment makes some test specialists quite uneasy. Human judgment is fallible. Competent judges, as has been implied, may disagree. What these specialists would like to find is a way of determining test validity that does not involve fallible human judgments. Empirical test validation might provide this way if it were not for the problem of finding a valid criterion. But the difficulties and uncertainties involved in getting directly valid criterion

measures are exactly as serious and troublesome as those of obtaining directly valid test scores. In fact, the two problems are almost identical, as Anastasi has indicated.¹⁵ There is no means of test validation that is completely empirical, is completely impersonal and objective, and avoids the vagueness and uncertainty of human judgments altogether.

To say that the exercise of judgment is unavoidable in constructing valid tests or determining their validity is not to say that one man's judgment is as good as the next or that any man's judgment is above rational and corrective criticism. The decisions of the test constructor reflect his knowledge and those values. The knowledge may be incomplete. The reasoning may be faulty. Even the values may reflect faulty knowledge or reasoning. Assumptions that have not been, but could be, tested may be involved in the test constructor's decisions.

There is no simple, uniform, wholly objective procedure for determining the validity of a test or of a test item. The judgment of the test constructor is inevitably involved in the process of test construction. But his judgment is not beyond criticism and improvement. It can be refined and given a sounder basis, as Gulliksen has argued persuasively.¹⁶ Not all the differences of opinion among test constructors with respect to questions of relevance are likely to be resolved quickly and surely by rational consideration and further experimental evidence. It is unlikely that a majority of the teachers of any subject will ever agree on one particular test as the ideal test of achievement in that subject. What can be expected is increasing awareness of, and agreement on, the essentials for validity in items purporting to test achievement in that subject. Capable, conscientious teachers are likely to make progressively better tests as they think and discuss what their tests ought to measure and as they invent and try out new tasks for their tests.¹⁷

TASKS AND TRAITS

This chapter has suggested that the validity of a test can be judged, in part at least, by studying the test itself. But does a test really measure what it appears to measure? Is this a test of scientific understanding,

¹⁵Anne Anastasi, "The Concept of Validity in the Interpretation of Test Scores," *Educational and Psychological Measurement*, vol. 10 (Spring 1950), 67-78.

¹⁶Gulliksen, "Intrinsic Validity,"

¹⁷Robert L. Ebel, "Obtaining and Reporting Evidence on Content Validity," *Educational and Psychological Measurement*, vol. 16 (1956), 269-82.

as its title claims, or is it really an intelligence test? Is that a test of intelligence, or is it really only a reading test?

If one is willing to accept as a definition of what a test is measuring a simple description of the tasks that the test requires a student to perform, then what a test appears to measure and what it really does measure will be practically identical. For example, the question, "What is the sum of $\frac{1}{2}$ and $\frac{1}{4}$?" appears to measure, and really does measure (beyond his ability to read), the student's ability to add two particular common fractions. The tasks in most classroom tests of educational achievement can be described as obviously and sensibly as this if one is willing to settle for an obvious, common-sense description.

But some are not. Some prefer to name their tests and describe what their tests are measuring, not in terms of the tasks they present, but in terms of the traits they presumably measure. So we have tests of intelligence, persistence, empathy, rigidity, creativity, anxiety, tolerance, perceptiveness, reasoning, and many other traits. For tests like these, the question of whether the test really measures what it claims to measure does arise, as indeed it should. Does the task of completing a figure analogy measure intelligence? Does ability to suggest unconventional uses for a brick measure creativity? Does ability to repeat a series of digits in reverse order measure memory?

Unfortunately, when such questions do arise, it is usually impossible to find satisfactory answers to them. For the traits these tests are supposed to measure are so highly generalized and so variable from one situation to another that a primary, intrinsically valid trait-defining test is seldom available for them. In the absence of such direct measures of the traits in question it is well nigh impossible to establish the validity of tests that claim to measure the traits indirectly.

Whenever measures of magnitude A , obtained by operations that define magnitude A , correlate with measures of magnitude B , obtained by operations that define magnitude B , then A measures B to some degree, and B measures A . But if we have no measures that define magnitude A , it is impossible to answer the question, "Do the operations we used to obtain measures B really measure A ?" For reasons like these we can not always tell for sure whether test B does in fact measure trait A .

The generality of meaning of many trait names and the variability of the behavior attributed to such traits, depending on the specific circumstances, have caused much trouble for those who try to construct tests of these traits. It is not uncommon to find that the scores on three different tests, all purporting to measure the same trait, show intercorrelations of near zero. Further, there is a dangerous tendency to over-generalize behavior predictions on the basis of trait names. Here is an example.

Applicants for positions on the police force in one city were shown this sign

Throw trash
in the
the trash cans.

and asked to comment on it. If they noticed the repetition of the word "the," they were given credit for perceptiveness and their chances of acceptance were improved.

All would agree that policemen need to be perceptive, to notice and remember important details like the license numbers of certain cars and descriptions of certain people. But there is little evidence, and really not much reason to expect, that people who are good at noticing printing errors in signs will also be good at noticing essential details at the scene of a crime.

Test items like the misprinted sign, which have only a verbal claim to direct relationship to the trait to be measured, do call for empirical validation. But perhaps they should be validated not as measures of perceptiveness, but as predictors of effectiveness on the police force. The trouble is that to do a good job of that kind of empirical validation is extremely difficult. The primary, directly valid measures of effectiveness are hard to design and harder still to use effectively. Fortunately, classroom tests can usually avoid most of these quicksands of trait measurement. If they are built carefully and skillfully enough they can have direct, primary validity.

VALIDITY AND RELIABILITY

The term "validity," as noted earlier in this chapter, is sometimes defined as the "accuracy with which the test measures what it is intended to measure." This is in contrast with "reliability," which can be defined as the "accuracy with which the test measures whatever it does measure." Clearly, a test that measures accurately what it is intended to measure also measures with equal or greater accuracy whatever it does measure. Hence to the degree that a test is valid it must also be reliable. Further, a test cannot measure what it is intended to measure more accurately than it measures whatever it *does* measure.¹⁸ Hence in order to be valid a test must be reliable.

¹⁸This statement is not contradicted by the fact that it is theoretically possible for a test to yield a validity coefficient that is numerically larger than its reliability coefficient for the same examinees. For the validity coefficient can reach its theoretical maximum only if the criterion for validity is a perfectly accurate measure of whatever the test does in fact measure. In that case the difference between what a test is intended to measure and what it does measure has disappeared. If error-free criterion measures are assumed for both reliability and validity, the index of validity can never be larger than the index of reliability.

The converse of the relation just stated between reliability and validity is not necessarily true, however. That is, a reliable test—one that measures accurately whatever it does measure—is not necessarily a valid test—one that measures accurately what it is intended to measure. Reliability is a necessary condition for validity, but it is not a sufficient condition.

Indeed, there is at least the possibility that a test constructor, working to improve the reliability of his test, might actually lower its validity. Some types of ability, like knowledge of word meanings or ability to solve numerical problems, are easier to measure reliably than other things, such as understanding of principles or ability to suggest appropriate action in a practical problem situation. If the test constructor concentrates on the types of ability that can be measured easily, to the neglect of others that also ought to be measured, his test could become more reliable but less valid.

Again, a homogeneous test is likely to be more reliable than a heterogeneous test. But if a course of instruction aims to develop a variety of different abilities or relatively unrelated understandings, an appropriate test of achievement for that course must show corresponding variety, or heterogeneity, in the items it includes. Making the test valid in this case may interfere somewhat with making it reliable.

The selection of items for high indices of discrimination, as pointed out in Chapter 14 on item analysis, tends to make the test more reliable. If the rejected items are low in discrimination because of low technical quality, the reliability is gained at no sacrifice in validity. But if they show low indices of discrimination because they are measuring something different from the majority of items in the test and if that something is an essential part of what the whole test ought to measure, then the elimination of those items from the test would increase reliability at the expense of validity.

None of these possibilities of opposition or interference between reliability and validity justifies any devaluation of high reliability as a goal in test construction. Reliability is essential to validity. The test constructor is not justified in taking shortcuts to reliability by measuring only those things that can be easily measured with high reliability or by making his test more homogeneous than is warranted by the complexity of what he is testing; neither is he justified in settling for an unreliable test. He must try to measure what needs to be measured as reliably as possible. If he does not, the validity of his test will suffer.

VALIDITY AND RELEVANCE

Reliability is clearly an important component of validity. Another component is relevance. A test that measures with a high degree of

accuracy what it is intended to measure is highly valid because it is both highly relevant and highly reliable. But a test could be highly reliable, in theory at least, without being highly valid if it lacked a high degree of relevance.

For almost all classroom tests of educational achievement, relevance is a matter of logical analysis and expert judgment. Unlike reliability, it ordinarily cannot be measured statistically, on the basis of experimental data, after the test has been given. Relevance must be built into the test.¹⁹ What a test actually does measure is determined by the test constructor as he works, step by step, to build the test. The cumulative effect of the decisions he makes at each step in the process determines the relevance of the test. If the individual items in the test are relevant, that is, if they require demonstration of mastery of some essential aspect of the course and if they sample proportionally all those essential aspects, then the test as a whole will be relevant.²⁰

Much of what this book has had to say about test construction has related to the problem of building relevance into the tests. This is especially true of the chapters on what can be measured, on test planning, and on the writing of true-false and multiple-choice items. But it may be useful to restate here some of the factors that are most likely to affect the relevance of classroom tests of educational achievement.

Chapter 3 suggested that the principal objective of all classroom instruction is to develop in the students a command of useful knowledge. To measure a student's achievement of this objective, a classroom test should present tasks that require a student to demonstrate his command of the knowledge, that is, his ability to use it. It should not reward sheer verbal memory. It should include novel, practical problems that require the application of knowledge. It should not present problems that can be solved by logical reasoning alone. The tasks in the test should require demonstration of achievement of the unique goals of instruction in the particular course. They should not be answerable successfully on the basis of general intelligence alone. The items in the test should be free of irrelevant clues that play into the hands of the sophisticated test taker. Above all, the tasks selected should relate to the most fundamental, the most central, and hence the most useful concepts and principles developed by the course.

The kind of validity we have been discussing, based on relevance and reliability, is the kind most classroom tests of educational achievement ought to have, and indeed the only kind they ordinarily can have. It is primary, direct, intrinsic validity. Sometimes it is referred to as "con-

¹⁹Edith M. Huddleson, "Test Development on the Basis of Content Validity," *Educational and Psychological Measurement*, vol. 16 (Autumn 1956), 283-93.

²⁰Roger T. Lennon, "Assumptions Underlying the Use of Content Validity," *Educational and Psychological Measurement*, vol. 16 (Autumn 1956), 294-304.

tent validity" or "curricular validity." How much of this kind of validity a test possesses cannot be determined by looking at the scores the test yields. Instead one must look at the test itself, at the rationale and specifications for the test, if they are available, and at directions for administering and scoring it.

CHARACTERISTICS OF VALIDITY

The validity of any test is clearly a matter of degree, not an all-or-none quality. Tests are not valid or invalid. They are more or less valid. Further, the validity of a test is not completely determined by the test itself. It depends on the purpose for which the test is used, the group with which it is used, and the way it is administered and scored. Instead of asking, "How valid is this *test*?" it would be more precise to ask, "How valid are the *scores* from this test when it is used in a specified way for a specified purpose with a specified group?" Most classroom tests are constructed with a very specific group and purpose in mind. Flaws in test administration might adversely affect the reliability of the scores but are not likely to affect relevance appreciably.

How long does the validity of a classroom test last? If copies of the test are returned to the students for discussion and remedial instruction after the test was given and if the students who take the same course in successive years have frequent contacts with each other, the validity of a test may be abraded rapidly. Or if the teacher begins to use test content for teaching purposes, that is, to teach for the test directly, then its validity will also suffer. The tasks in any test are only a sample. At best they are a sample of a population of tasks any one of which the student would meet for the first time were it to be included in the test. Direct teaching for any of the tasks in a test tends to spoil the test as a representative sample from a population of novel tasks.

On the other hand, where no extensive use is made of the test for teaching and where contacts between members of successive classes are likely to be few and casual, the same test may retain its validity for a long time. Good tests are hard enough to build to make a teacher or professor consider carefully the pros and cons of using test items for teaching purposes.

HANDOUT 7

ITEM ANALYSIS

Obviously some test items are more difficult than others and some have more discriminating power than others. Discriminating power is the ability of a test item to discriminate between those who performed well on the test as compared to those who did poorly. To conduct the item analysis the instructor could use the following steps:

- a. Rank and sort the papers by score with the highest score on top and the lowest score on the bottom.
- b. Separate the stack of papers into three piles: the top 25% into one stack, the middle 50% into a second stack, and the bottom 25% into a third stack. Determine T, the number of papers in the top 25% stack (which is the same as the number of papers in the bottom 25% stack).
- c. Count or tally the number of people in the high group and the low group who got each item correct. For each item, therefore, you should have a number which we will call H (number who got the item correct in the high group) and another number L (number who got the item correct in the low group).
- d. Calculate item difficulty (D)

$$D = (H + L) / 2T$$

Items of approximately 0.50 difficulty will generally have the greatest discrimination power.

- e. Calculate Discrimination Index, (D.I.)

$$D.I. = (H - L) / T$$

If all the people in the high group and none of the people in the low group got the item correct, $H - L = H$ and $D.I. = 1.0$. This would be a perfectly discriminating item and would be a very rare occurrence. In actual practice, the D.I. will generally be much less than 1.0. In general, questions with discrimination indices below 0.2 are considered poor discriminators while those greater than 0.5 are good discriminators. You will probably want to have a few items that everyone answers correctly, so some items will need to be poor discriminators.

- f. Record the item difficulty (D) and Discrimination Index (D.I.) on the back of the item card along with the date of the test when the item was used.

HANDOUT 8

NORM VS. CRITERION-REFERENCED GRADING

From: Mehrens, William A., and Lehmann, Irvin J., *Measurement and Evaluation in Education and Psychology*, Holt, Rinehart and Winston, Inc., New York, 1973, pp. 63-77.

NORM- AND CRITERION-REFERENCED MEASUREMENT

There has been considerable discussion about educational measurement with regard to the distinctions between, and the advantages and disadvantages of, norm-referenced and criterion-referenced measurement. The basic issue concerns interpretation, or how we derive meaning from a score, but related to this is a discussion of different purposes of measurement and techniques of constructing tests.

The distinction between the two types of scores seems clear enough. If we interpret a score of an individual by comparing his score to those of other individuals (called a norm group) this would be norm referencing. If we interpret a person's performance by comparing it to some specified behavioral criterion of proficiency, this would be criterion referencing. To polarize the distinction, we could say that the focus of a normative score is on how many of Johnny's peers do not perform (score) as well as he does; the focus of a criterion-referenced score is on what it is that Johnny can do. Of course we can, and often do, interpret a single test score both ways. In norm referencing we might make a statement that "John did better than 80 percent of the students in a test on addition of whole numbers." In criterion referencing we might say that "John got 70 percent of the items correct in a test on addition of whole numbers." Ebel (1962, p. 19) pointed out that "to be meaningful any test scores must be related to test content *as well as* to the scores of other examinees." Jackson (1970, p. 2) agreed and correctly stated that any test will sample the content of some specified domain and that there is always an implicit behavioral element. However, in norm-referenced measurement, in contrast to criterion-referenced measurement, "the inference is of the form—'more (or less) of trait *x* than the mean amount in population *y*'—rather than some specified amount that is meaningful in isolation" (Jackson, 1970, p. 2).

Depending on the decision we wish to make, one of these two ways of interpreting the score will likely be more useful than the other, and it has been suggested by many that for those situations where criterion referencing is likely to be the most useful, a measuring instrument should be constructed and analyzed differently than what "classical measurement theory" suggests.³

Classical measurement theory originated and developed from an interest in measuring the aptitudes of individuals. Thus, the emphasis has been on devising measures that differentiate or discriminate among individuals at all points along the continuum. That is, a test is devised so that individuals who have more aptitude will score higher on the test than those who possess less. The interpretation here of "how much" aptitude a person possesses is a normative one. The same test theory, and this tendency to construct tests that discriminate among individuals, has carried over to achievement testing. There are times, however, when it is not necessary, and perhaps not even advisable, to differentiate individuals at all degrees of achievement. We may simply want to find out whether individuals have achieved a specific set of objectives. In other words, we reference a person's score against a criterion. Thus, there are really two different goals or objectives in *achievement testing*: (1) To discriminate among all individuals according to their degrees of achievement, and (2) to determine whether a person has achieved a specific set of objectives.⁴ Textbooks on constructing achievement tests have typically discussed testing techniques that lead to the first goal. Some educators feel that this is somewhat deplorable, and suggest that more emphasis be placed on achieving the second goal.

We will briefly discuss the uses of normative and criterion-referenced measurement in this section. We will focus on mastery testing as the most popular example of criterion-referenced measurement. We will introduce some of the differences in constructing the two types of measures, but will expand this discussion in later sections of the book on reliability, validity, and item analysis.

Uses for Criterion-Referenced Measurement

The recent support for criterion-referenced measurement seems to have originated in large part from the emphases on behavioral objectives, the sequencing and individualization of instruction, the development of programmed materials, a learning theory that *suggests* that most anybody can learn most anything if given enough time, and a belief that norm referencing promotes unhealthy competition and is injurious to low-scoring students' self-concepts.

³ In fact, some writers define criterion-referenced tests as those deliberately constructed so as to yield scores directly interpretable in terms of specified performance standards. (See Glasser and Nitko, 1971, and Jackson, 1970.)

⁴ Of course one could discriminate with a criterion-referenced measure, but that is not the primary purpose of such a measure.

If we can specify important objectives in behavioral terms, then, many would argue, the important consideration is to ascertain whether a student obtained those objectives rather than his position relative to other students. Traditionally, the principal use of criterion-referenced measurement has been in "mastery tests." A mastery test is a particular type of criterion-referenced test. Mastery, as the word is typically used, connotes an either/or situation. The person has either achieved (mastered) the objective(s) satisfactorily or he has not. Criterion-referenced testing in general could also apply to degrees of performance. Mastery tests are used in programs of individualized instruction, such as the Individually Prescribed Instruction (IPI) program (Lindvall and Bolvin, 1967), or in the mastery learning model devised by Bloom (1968). Such instructional programs are composed of units or modules, usually considered hierarchical, each based on one or more instructional objectives. Each individual is required to work on the unit until he has achieved a specified minimum level of achievement. Thus, he is considered to have "mastered" the unit. In such programs, the instructional decision of what to do with a student is *not* dependent on how his performance compares to others. If he has performed adequately on the objectives, then the decision is to move on to the next unit of study. If he has not, then he is required to restudy the material (although perhaps using a different procedure) covered in the test until he performs adequately, that is, "masters" the material. If instructional procedures are organized so that time is the dimension that varies and degree of mastery is held constant, then mastery tests should be used in greater proportion than they are now.

Employing the individually prescribed instruction or mastery model of learning is not the only use for criterion-referenced measures. One may also use such data to evaluate (make decisions about) instructional programs. In order to determine whether specific instructional treatments or procedures have been successful, it is necessary to have data about the outcomes on the specific objectives the program was designed to teach. A measure comparing students to each other (norm referencing) may not give so effective data as a measure comparing each student's performance to the objectives.

Also, criterion-referenced measures offer certain benefits for instructional decision making within the classroom. The diagnosis of specific difficulties, accompanied by a prescription of certain instructional treatments, is necessary in instruction whether or not one uses a mastery approach to learning.

Mastery Testing

Since mastery testing is apparently the most frequently used criterion-referenced approach and it is being so strongly advocated as "one of the most powerful ideas beginning to shape educational views and practices" (Block, 1971, iii), we will discuss the rationale behind it a bit more thoroughly. The idea of mastery learning and mastery testing is certainly not new (see Washburne, 1922; Morrison, 1926). But, as Baker (1971, p. 65) suggested, "A

considerable literature relating to the evils of mastery tests exists, and much of the work of early educational psychologists was in reaction to the unreal requirement that all pupils achieve criterion performance." The basic idea of mastery learning, however, was revitalized with the publication of a paper by Carroll (1963) entitled "A Model of School Learning." Essentially, the model suggests that the degree of learning is a function of the time the student spends on the material, divided by the time needed. More precisely, Carroll suggested that the degree of learning is some function of the time allowed and the perseverance of the student, divided by the student's aptitude for the task, his ability to understand the instruction, and the quality of instruction.

Bloom (1968) agreed with the basics of the model and suggested that the degree of learning required should be fixed at some "mastery" level and that the instructional variables should be manipulated so that all (or almost all) students achieve mastery. Bloom stated that "Most students (perhaps over 90 percent) can master what we have to teach them" (Bloom, 1968, p. 1). (Interestingly enough, this bold statement was made prior to defining mastery, for later in the same paragraph Bloom stated that "Our basic task is to determine what we mean by mastery. . . ." How do we know 90 percent can master what we have to teach them if we haven't defined what mastery is?) If the model is correct and if people should all persevere until they have "mastered" the material, then the mastery learning model of instruction should be employed—and mastery testing needs to be used to determine whether mastery has occurred.

Tentative evidence (Block, 1971) suggests that in many subject-matter areas all students can achieve some (often arbitrarily) defined level of mastery, although—as Carroll (1971, p. 31) pointed out—if the task is very difficult or depends upon special aptitudes, there may be a number of students who never make it. Becoming a 4-minute miler or a concert pianist are examples.

Excluding the extreme 5 percent of the students, the time ratio between slower and faster students is about 6 to 1, although Bloom et al. (1971, p. 51) suggested that this may be reduced to about 3 to 1. Glaser (1968, p. 28) reported that in three years of Individually Prescribed Instruction in mathematics, one student had covered 73 units, one only 13. The mean was 37 units; the standard deviation, 12. Whether or not it is worthwhile educationally to have a student persist for three to six days, weeks, or years on a task others can complete in one day, week, or year is more debatable. They probably should for those few basic skills that indeed are essential for survival and which must be achieved before further hierarchical skills can be acquired. For other things we attempt to teach in school—like understanding of modern literature—a mastery model should probably not be employed. There is even some doubt if it would work for such a subject. As both Block (1971, p. 66) and Bloom (1971b, p. 33) pointed out, mastery learning strategies

are more effective for *closed* subjects (those whose content has not changed for some time) and those that emphasize convergent rather than divergent thinking. The implications for education of this admission by mastery advocates are not always fully appreciated. However, Cronbach, in a reaction to one of Bloom's mastery learning papers, brought the issue into sharp focus:

I find the concept of mastery severely limiting, and in trying to find out where my distress lies, I finally focused on one word in the Bloom paper: he states that mastery learning is *closed*. Training is closed. In education the problems are open. . . .

I see educational development as continuous and open-ended. "Mastery" seems to imply that at some point we get to the end of what is to be taught. (Cronbach, 1971, pp. 52, 53.)

Mastery Testing and Grades Bloom began his 1968 paper with the statement: "Each teacher begins a new term with the expectation that about a third of his students will adequately learn what he has to teach." As his paper proceeded, it is apparent that he based this statement on the fact that if a teacher graded on a normal curve (which many do not) about 33 percent would receive A's or B's. He suggested that "we proceed in our teaching as though only the minority of our students should be able to learn what we have to teach." What Bloom did, of course, was to attach his own absolute standards of interpretation to a normative reporting system. If a teacher gives a student a C because he (the teacher) is using a normative system and because the student is average (that is, in comparison to others, he was close to the middle), the grade of C in no way indicates inadequate performance. The symbols A through F can obviously take on any meaning a person assigns them. But if a teacher assigns them on a relative basis, it is hardly appropriate for someone else to interpret them on an absolute basis. We do not doubt that if those who score below a normative-based A were to continue studying the course material, many of them could in time achieve as well as the A students did originally. But we wonder why average performance should be interpreted as inadequate by a student, teacher, or educational psychologist, or why mastery should, as Bloom suggested, be considered identical to an A grade. One may well want students to "master" a certain amount of material before being allowed to advance to a higher unit or course. Whether one labels this mastery as an A or a D is irrelevant as long as the meaning is made clear. We will talk more about his point in Chapter 17.

Mastery Learning and Self-Concept Bloom et al. felt that:

Mastery learning can be one of the more powerful sources of mental health. We are convinced that many of the neurotic symptoms displayed by high school and college students are exacerbated by painful and frustrating experiences in school learning. If 90 percent of the students are given positive indications of adequacy in learning, one might expect them to need less and less in the way of emotional therapy and psychological help. Conversely,

frequent indications of failure and learning inadequacy are bound to be accompanied by increased self-doubt in the student. (Bloom et al., 1971, p. 56.)

With the exception of the first sentence in the preceding statement, we are inclined to agree (although no good evidence exists). The question remains, however, whether the mastery learning approach will decrease a student's painful and frustrating experiences in school learning, or whether it will increase the positive indications of adequacy in learning. We are inclined to think that the student who is forced to spend ten weeks on a unit that another student finished in two weeks will perceive that experience to be more painful and that he will have experienced more indications of failure and inadequacy than if we had assigned him a grade of C or D after the first two weeks. Contrary to Bloom's impression, many students perceive C's and D's to be quite adequate. Of course Bloom's hope is that the time ratio may be reduced to a vanishing point (Bloom, 1971a, p. 8), but he admits such a point has never been reached, and it is indeed very doubtful that individual differences in school will or should reach a vanishing point.

These last few paragraphs have not been written to denigrate the mastery learning approach. The mastery learning-testing model is useful for some situations. But it is probably not the pervasive good that some of its proselytizers would suggest.

Uses for Norm-Referenced Measurement

As stated in the beginning of this section, most testing and the theory of testing have been based on the norm-referenced approach. There is little argument that such an approach is useful in aptitude testing where we wish to make differential predictions. It is also often very useful to achievement testing. Many educators would agree with Gronlund's (1971, p. 139) statement: "In measuring the extent to which pupils are achieving our course objectives, we have no absolute standard by which to determine their progress. A pupil's achievement can be regarded as high or low only by comparing it with the achievement of other pupils."

Accepting this view, the role of a measuring device is to give us as reliable a rank ordering of the pupils, with respect to the achievement we are measuring, as possible. Knowing what we do about individual differences, it is obvious that students will learn differing amounts of subject matter even under a mastery learning approach. It may be that all students, or at least a high percentage of them, have learned a significant enough portion of a teacher's objectives to be categorized as having "mastered" the essentials of the course or unit. But some of these students have learned more than others, and it seems worthwhile to employ measurement techniques that identify these pupils. In the first place, students want and deserve recognition for accomplishment that goes beyond the minimum. If we would continually give

only mastery tests, those students who accomplish at a higher level would lose one of the important extrinsic rewards of learning, that is, recognition for such accomplishments.

Perhaps a more important reason for discrimination testing than student recognition is in its benefits for decision making. If two physicians have mastered surgery, but one has mastered it better, which one do you wish to have operate on you? (For that matter, even if two physicians had equally mastered their training program, one would probably want some norm-referencing information about time to completion. If one physician is such a slow learner that it takes him five times as long to learn the material as the other one, it is probably safe to assume that after he has been on the job ten years, he will not be so up to date on current medical practices as the fast learner.) If two teachers have mastered the basics of teaching, but one is a much better teacher, which do we want to hire? If two students have mastered first-semester algebra, but one has learned it much better (or faster, time being norm-referenced), which should receive the most encouragement to continue in mathematics? We probably all agree on the answers to these questions. However, if we have not employed evaluation techniques that allow us to differentiate between the individuals, we cannot make these types of decisions. Certainly, norm-referenced measures are the most helpful in fixed-quota selection decisions. For example, if there are a limited number of openings in a pilot training school, the school would want to select the best of the applicants—even though all may be above some "mastery level."

Comparing the Two Measures

The decision concerning when to use norm-referenced and when to use criterion-referenced measures depends upon the kind of decision one wishes to make. In achievement testing, it depends mostly upon the instructional procedures employed. If instruction is structured so that time is the variable, and a student keeps at a task until he has mastered it, then we should use mastery testing. This type of instruction is often employed in individualized instruction. If instruction is structured so that time of exposure is constant, then students will achieve at different levels, and we should attempt to discern this differential achievement with a test that discriminates, although we might well want to attach both normative and criterion-referenced meaning to the score. Which instructional procedure should be used depends upon the structure and importance of the subject matter being taught.

There surely are some subjects so structured that it is futile to teach higher concepts until basic ones have been mastered. For example, a student cannot do long division until he can subtract and multiply at some basic level (although precisely what that level is is unknown). This is certainly not true of all subjects, however. We do not really need to have mastered (or even have read) *A Tale of Two Cities* before reading *Catcher in the Rye*, or vice versa.

Likewise, there may well be some skills or knowledges so important that all students should master them, regardless of how long it takes. Knowing how to spell one's name probably fits in this category. But, again, this is not true of all subjects. With regard to this point, Ebel stated:

We might be willing to allow one student a week to learn what another can learn in a day. But sum these differences over the myriads of things to be learned. Does anyone, student, teacher, or society, want to see one person spend 16 or 24 years getting the same elementary education another can get in eight? Should it be those least able to learn quickly who spend the largest portion of their lives in trying to learn? Our present practice is quite the reverse. Those who are facile in learning make a career of it. Those who are not find other avenues of service, fulfillment and success. (Ebel, 1969, p. 12.)

Gronlund (1970, p. 32) made a distinction between objectives that are considered as *minimum essentials* and those that encourage *maximum development*. For the former, one would want to employ mastery testing; for the latter, discrimination testing.

Thus, we are suggesting that there is a place for both mastery (criterion-referenced) and discrimination (norm-referenced) testing. The way most schools are currently organized, with time of instruction constant for all individuals and degree of learning the variable, discrimination testing should be prevalent. However, as more individualized instructional processes are used and as more is learned about how various subject matters should be sequenced, mastery testing may increase in importance. Mastery testing is probably more important in the early elementary grades than later in school. Table 3-1 is an oversimplification, but it summarizes the polar positions.

Constructing Criterion- and Norm-Referenced Achievement Tests

As mentioned earlier, traditional test theory and techniques of test construction have been developed on the assumption that the purpose of a test is to discriminate among individuals. If the purpose of a test is to compare each individual to a standard, then it is irrelevant whether or not the individuals differ from each other. Thus, some of the criteria of a measuring instrument considered essential for a norm-referenced measure are not important for criterion-referenced measures. (See Popham and Husek, 1969.) What one looks for in item analysis, reliability, and some types of validity are different in a criterion-referenced measure.

However, for many aspects of test construction such as considering the objectives, preparing test blueprints, and actually wording the items, there are more similarities than differences in the preparation of normative and criterion-referenced tests. There is a healthy emphasis, for example, on making

**TABLE 3-1 Summary Position Statements on
the Mastery-Discrimination Controversy**

Advocates of discrimination testing believe that—	Advocates of mastery testing believe that—
1. Mastery can be only arbitrarily defined and that there are important decisions to be reached from knowledge of degrees of achievement. 2. Most subject matter is not so structured that understanding one concept requires "complete mastery" of a previous concept. 3. Varying time rather than amount learned is not the most beneficial or efficient instructional strategy. 4. All students need not, and probably should not, learn many (if any) things to the same degree of proficiency. 5. Even if item 4 were a reasonable goal, what we know about individual differences suggests that even by making time a variable, student proficiencies are not likely to be very homogeneous. 6. If differential learning exists, it is wasteful not to measure it, since it is a potential aid in decision making.	1. Mastery is definable in a meaningful, useful way, and degrees of mastery for many subjects either do not exist or are unimportant. 2. Most subject matter is hierarchically structured. 3. For hierarchically structured subjects it is an unwise pedagogical decision to attempt to teach higher concepts if the basics have not been mastered. 4. Therefore, it is better to allow variations in the time spent in learning than in the amount learned (that is, it is better for a student to learn fewer things more thoroughly). 5. All pupils enrolled in a course should learn the same things from that course.

sure that items measure certain specified objectives in criterion-referenced tests. But this same characteristic should hold for a norm-referenced instrument.

Differences and similarities between the two approaches—as they pertain to reliability, validity, test blueprints, item writing, and item analysis—will all be discussed in later sections of the text.